
JMIR AI

Volume 3 (2024) ISSN 2817-1705 Editors-in-Chief: Khaled El Emam, Bradley Malin, PhD

Contents

Tutorial

- Generating Synthetic Electronic Health Record Data Using Generative Adversarial Networks: Tutorial
([e52615](#))
Chao Yan, Ziqi Zhang, Steve Nyemba, Zhuohang Li. 7

Original Papers

- Cost, Usability, Credibility, Fairness, Accountability, Transparency, and Explainability Framework for Safe and Effective Large Language Models in Medical Education: Narrative Review and Qualitative Study
([e51834](#))
Majdi Quttainah, Vinaytosh Mishra, Somayya Madakam, Yotam Lurie, Shlomo Mark. 21
- Privacy-Preserving Federated Survival Support Vector Machines for Cross-Institutional Time-To-Event Analysis: Algorithm Development and Validation ([e47652](#))
Julian Späth, Zeno Sewald, Niklas Probul, Magali Berland, Mathieu Almeida, Nicolas Pons, Emmanuelle Le Chatelier, Pere Ginès, Cristina Solé, Adrià Juanola, Josch Pauling, Jan Baumbach. 107
- Machine Learning–Based Prediction for High Health Care Utilizers by Using a Multi-Institutional Diabetes Registry: Model Training and Evaluation ([e58463](#))
Joshua Tan, Le Quan, Nur Salim, Jen Tan, Su-Yen Goh, Julian Thumboo, Yong Bee. 119
- Predictive Performance of Machine Learning–Based Models for Poststroke Clinical Outcomes in Comparison With Conventional Prognostic Scores: Multicenter, Hospital-Based Observational Study ([e46840](#))
Fumi Irie, Koutarou Matsumoto, Ryu Matsuo, Yasunobu Nohara, Yoshinobu Wakisaka, Tetsuro Ago, Naoki Nakashima, Takanari Kitazono, Masahiro Kamouchi. 143
- Improving Risk Prediction of Methicillin-Resistant Staphylococcus aureus Using Machine Learning Methods With Network Features: Retrospective Development Study ([e48067](#))
Methun Kamruzzaman, Jack Heavey, Alexander Song, Matthew Bielskas, Parantapa Bhattacharya, Gregory Madden, Eili Klein, Xinwei Deng, Anil Vullikanti. 163
- Identifying Symptoms Prior to Pancreatic Ductal Adenocarcinoma Diagnosis in Real-World Care Settings: Natural Language Processing Approach ([e51240](#))
Fagen Xie, Jenny Chang, Tiffany Luong, Bechien Wu, Eva Lustigova, Eva Shrader, Wansu Chen. 181
- Sepsis Prediction at Emergency Department Triage Using Natural Language Processing: Retrospective Cohort Study ([e49784](#))
Felix Brann, Nicholas Sterling, Stephanie Frisch, Justin Schrager. 192

Learning From International Comparators of National Medical Imaging Initiatives for AI Development: Multiphase Qualitative Study ([e51168](#))
 Kassandra Karpathakis, Emma Pencheon, Dominic Cushnan. 203

Framework for Ranking Machine Learning Predictions of Limited, Multimodal, and Longitudinal Behavioral Passive Sensing Data: Combining User-Agnostic and Personalized Modeling ([e47805](#))
 Tahsin Mullick, Sam Shaaban, Ana Radovic, Afsaneh Doryab. 214

Predicting Workers' Stress: Application of a High-Performance Algorithm Using Working-Style Characteristics ([e55840](#))
 Hiroki Iwamoto, Saki Nakano, Ryotaro Tajima, Ryo Kiguchi, Yuki Yoshida, Yoshitake Kitanishi, Yasunori Aoki. 237

Optimizing Clinical Trial Eligibility Design Using Natural Language Processing Models and Real-World Data: Algorithm Development and Validation ([e50800](#))
 Kyeryoung Lee, Zongzhi Liu, Yun Mai, Tomi Jun, Meng Ma, Tongyu Wang, Lei Ai, Ediz Calay, William Oh, Gustavo Stolovitzky, Eric Schadt, Xiaoyan Wang. 248

A Comparison of Personalized and Generalized Approaches to Emotion Recognition Using Consumer Wearable Devices: Machine Learning Study ([e52171](#))
 Joe Li, Peter Washington. 270

Sample Size Considerations for Fine-Tuning Large Language Models for Named Entity Recognition Tasks: Methodological Study ([e52095](#))
 Zoltan Majdik, S Graham, Jade Shiva Edward, Sabrina Rodriguez, Martha Karnes, Jared Jensen, Joshua Barbour, Justin Rousseau. 282

Augmenting Telepostpartum Care With Vision-Based Detection of Breastfeeding-Related Conditions: Algorithm Development and Validation ([e54798](#))
 Jessica De Souza, Varun Viswanath, Jessica Echterhoff, Kristina Chamberlain, Edward Wang. 294

Enhancing Type 2 Diabetes Treatment Decisions With Interpretable Machine Learning Models for Predicting Hemoglobin A1c Changes: Machine Learning Model Development ([e56700](#))
 Hisashi Kurasawa, Kayo Waki, Tomohisa Seki, Akihiro Chiba, Akinori Fujino, Katsuyoshi Hayashi, Eri Nakahara, Tsuneyuki Haga, Takashi Noguchi, Kazuhiko Ohe. 314

Identifying Frailty in Older Adults Receiving Home Care Assessment Using Machine Learning: Longitudinal Observational Study on the Role of Classifier, Feature Selection, and Sample Size ([e44185](#))
 Cheng Pan, Hao Luo, Gary Cheung, Huiquan Zhou, Reynold Cheng, Sarah Cullum, Chuan Wu. 331

Development of Lung Cancer Risk Prediction Machine Learning Models for Equitable Learning Health System: Retrospective Study ([e56590](#))
 Anjun Chen, Erman Wu, Ran Huang, Bairong Shen, Ruobing Han, Jian Wen, Zhiyong Zhang, Qinghua Li. 343

Mitigating Sociodemographic Bias in Opioid Use Disorder Prediction: Fairness-Aware Machine Learning Framework ([e55820](#))
 Mohammad Yaseliani, Md Noor-E-Alam, Md Hasan. 354

Traditional Machine Learning, Deep Learning, and BERT (Large Language Model) Approaches for Predicting Hospitalizations From Nurse Triage Notes: Comparative Evaluation of Resource Management ([e52190](#))
 Dhavalkumar Patel, Prem Timsina, Larisa Gorenstein, Benjamin Glicksberg, Ganesh Raut, Satya Cheetirala, Fabio Santana, Jules Tamegue, Arash Kia, Eyal Zimlichman, Matthew Levin, Robert Freeman, Eyal Klang. 380

Obtaining the Most Accurate, Explainable Model for Predicting Chronic Obstructive Pulmonary Disease: Triangulation of Multiple Linear Regression and Machine Learning Methods ([e58455](#))
 Arnold Kamis, Nidhi Gadia, Zilin Luo, Shu Ng, Mansi Thumbar. 391

Online Health Search Via Multidimensional Information Quality Assessment Based on Deep Language Models: Algorithm Development and Validation (e42630) Boya Zhang, Nona Naderi, Rahul Mishra, Douglas Teodoro.	409
Multiscale Bowel Sound Event Spotting in Highly Imbalanced Wearable Monitoring Data: Algorithm Development and Validation Study (e51118) Annalisa Baronetto, Luisa Graf, Sarah Fischer, Markus Neurath, Oliver Amft.	425
Leveraging Machine Learning to Develop Digital Engagement Phenotypes of Users in a Digital Diabetes Prevention Program: Evaluation Study (e47122) Danissa Rodriguez, Ji Chen, Ratnalekha Viswanadham, Katharine Lawrence, Devin Mann.	441
Behavioral Nudging With Generative AI for Content Development in SMS Health Care Interventions: Case Study (e52974) Rachel Harrison, Ekaterina Lapteva, Anton Bibin.	454
Identifying Links Between Productivity and Biobehavioral Rhythms Modeled From Multimodal Sensor Streams: Exploratory Quantitative Study (e47194) Runze Yan, Xinwen Liu, Janine Dutcher, Michael Tumminia, Daniella Villalba, Sheldon Cohen, John Creswell, Kasey Creswell, Jennifer Mankoff, Anind Dey, Afsaneh Doryab.	473
Impact of a Digital Scribe System on Clinical Documentation Time and Quality: Usability Study (e60020) Marieke van Buchem, Ilse Kant, Liza King, Jacqueline Kazmaier, Ewout Steyerberg, Martijn Bauer.	487
Machine Learning Methods Using Artificial Intelligence Deployed on Electronic Health Record Data for Identification and Referral of At-Risk Patients From Primary Care Physicians to Eye Care Specialists: Retrospective, Case-Controlled Study (e48295) Joshua Young, Chin-Wen Chang, Charles Scales, Saurabh Menon, Chantal Holy, Caroline Blackie.	502
Risk Perception, Acceptance, and Trust of Using AI in Gastroenterology Practice in the Asia-Pacific Region: Web-Based Survey Study (e50525) Wilson Goh, Kendrick Chia, Max Cheung, Kalya Kee, May Lwin, Peter Schulz, Minhu Chen, Kaichun Wu, Simon Ng, Rashid Lui, Tiing Ang, Khay Yeoh, Han-mo Chiu, Deng-chyang Wu, Joseph Sung.	518
Assessment of ChatGPT-3.5's Knowledge in Oncology: Comparative Study with ASCO-SEP Benchmarks (e50442) Roupen Odabashian, Donald Bastin, Georden Jones, Maria Manzoor, Sina Tangestaniapour, Malke Assad, Sunita Lakhani, Maritsa Odabashian, Sharon McGee.	531
The Impact of Expectation Management and Model Transparency on Radiologists' Trust and Utilization of AI Recommendations for Lung Nodule Assessment on Computed Tomography: Simulated Use Study (e52211) Lotte Ewals, Lynn Heesterbeek, Bin Yu, Kasper van der Wulp, Dimitrios Mavroeidis, Mathias Funk, Chris Snijders, Igor Jacobs, Joost Nederend, Jon Pluyter, e/MTIC Oncology group.	539
Evaluation of Generative Language Models in Personalizing Medical Information: Instrument Validation Study (e54371) Aidin Spina, Saman Andalib, Daniel Flores, Rishi Vermani, Faris Halaseh, Ariana Nelson.	556
Perceptions of Family Physicians About Applying AI in Primary Health Care: Case Study From a Premier Health Care Organization (e40781) Muhammad Waheed, Lu Liu.	573
Feasibility of Multimodal Artificial Intelligence Using GPT-4 Vision for the Classification of Middle Ear Disease: Qualitative Study and Validation (e58342) Masao Noda, Hidekane Yoshimura, Takuya Okubo, Ryota Koshu, Yuki Uchiyama, Akihiro Nomura, Makoto Ito, Yutaka Takumi.	590

Evaluating the Diagnostic Performance of Symptom Checkers: Clinical Vignette Study (e46875) Mohammad Hammoud, Shahd Douglas, Mohamad Darmach, Sara Alawneh, Swapnendu Sanyal, Youssef Kanbour.	601
Near Real-Time Syndromic Surveillance of Emergency Department Triage Texts Using Natural Language Processing: Case Study in Febrile Convulsion Detection (e54449) Sedigh Khademi, Christopher Palmer, Muhammad Javed, Gerardo Dimaguila, Hazel Clothier, Jim Buttery, Jim Black.	619
Enhancing Clinical Relevance of Pretrained Language Models Through Integration of External Knowledge: Case Study on Cardiovascular Diagnosis From Electronic Health Records (e56932) Qiu hao Lu, Andrew Wen, Thien Nguyen, Hongfang Liu.	633
Health Care Professionals' and Parents' Perspectives on the Use of AI for Pain Monitoring in the Neonatal Intensive Care Unit: Multisite Qualitative Study (e51535) Nicole Racine, Cheryl Chow, Lojain Hamwi, Oana Bucsea, Carol Cheng, Hang Du, Lorenzo Fabrizi, Sara Jasim, Lesley Johannsson, Laura Jones, Maria Laudiano-Dray, Judith Meek, Neelum Mistry, Vibhuti Shah, Ian Stedman, Xiaogang Wang, Rebecca Riddell.	645
Comparing the Efficacy and Efficiency of Human and Generative AI: Qualitative Thematic Analyses (e54482) Maximo Prescott, Samantha Yeager, Lillian Ham, Carlos Rivera Saldana, Vanessa Serrano, Joey Narez, Dafna Paltin, Jorge Delgado, David Moore, Jessica Montoya.	657
Reidentification of Participants in Shared Clinical Data Sets: Experimental Study (e52054) Daniela Wiewert, Bradley Malin, Joseph Duffy, Rene Utianski, John Stricker, David Jones, Hugo Botha.	670
Predictive Modeling of Hypertension-Related Postpartum Readmission: Retrospective Cohort Analysis (e48588) Jinxin Tao, Ramsey Larson, Yonatan Mintz, Oguzhan Alagoz, Kara Hoppe.	689
Identifying Patterns of Smoking Cessation App Feature Use That Predict Successful Quitting: Secondary Analysis of Experimental Data Leveraging Machine Learning (e51756) Leeann Siegel, Kara Wiseman, Alex Budenz, Yvonne Prutzman.	702
Use of Deep Neural Networks to Predict Obesity With Short Audio Recordings: Development and Usability Study (e54885) Jingyi Huang, Peiqi Guo, Sheng Zhang, Mengmeng Ji, Ruopeng An.	714
Evaluating Literature Reviews Conducted by Humans Versus ChatGPT: Comparative Study (e56537) Mehrnaz Mostafapour, Jacqueline Fortier, Karen Pacheco, Heather Murray, Gary Garber.	721
Leveraging Temporal Trends for Training Contextual Word Embeddings to Address Bias in Biomedical Applications: Development Study (e49546) Shunit Agmon, Uriel Singer, Kira Radinsky.	731
Understanding the Long Haulers of COVID-19: Mixed Methods Analysis of YouTube Content (e54501) Alexis Jordan, Albert Park.	744
Role of Ethics in Developing AI-Based Applications in Medicine: Insights From Expert Interviews and Discussion of Implications (e51204) Lukas Weidener, Michael Fischer.	763
Beyond the Hype—The Actual Role and Risks of AI in Today's Medical Practice: Comparative-Approach Study (e49082) Steffan Hansen, Carl Brandt, Jens Søndergaard.	778

An Environmental Uncertainty Perception Framework for Misinformation Detection and Spread Prediction in the COVID-19 Pandemic: Artificial Intelligence Approach ([e47240](#))
 Jiahui Lu, Huibin Zhang, Yi Xiao, Yingyu Wang. 786

The “Magical Theory” of AI in Medicine: Thematic Narrative Analysis ([e49795](#))
 Giorgia Lorenzini, Laura Arbelaez Ossa, Stephen Milford, Bernice Elger, David Shaw, Eva De Clercq. 804

Reviews

Approaches for the Use of AI in Workplace Health Promotion and Prevention: Systematic Scoping Review ([e53506](#))
 Martin Lange, Alexandra Löwe, Ina Kayser, Andrea Schaller. 35

Exploring Machine Learning Applications in Pediatric Asthma Management: Scoping Review ([e57983](#))
 Tanvi Ojha, Atushi Patel, Krishihan Sivapragasam, Radha Sharma, Tina Vosoughi, Becky Skidmore, Andrew Pinto, Banafshe Hosseini. 49

Viewpoints

Regulatory Frameworks for AI-Enabled Medical Device Software in China: Comparative Analysis and Review of Implications for Global Manufacturer ([e46871](#))
 Yu Han, Aaron Ceros, Jeroen Bergmann. 65

Toward Clinical Generative AI: Conceptual Framework ([e55957](#))
 Nicola Bragazzi, Sergio Garbarino. 76

The Utility and Implications of Ambient Scribes in Primary Care ([e57673](#))
 Puneet Seth, Romina Carretas, Frank Rudzicz. 93

The Dual Nature of AI in Information Dissemination: Ethical Considerations ([e53505](#))
 Federico Germani, Giovanni Spitale, Nikola Biller-Andorno. 97

Research Letters

Can Large Language Models Replace Therapists? Evaluating Performance at Simple Cognitive Behavioral Therapy Tasks ([e52500](#))
 Nathan Hodson, Simon Williamson. 135

What Is the Performance of ChatGPT in Determining the Gender of Individuals Based on Their First and Last Names? ([e53656](#))
 Paul Sebo. 498

Corrigenda and Addendas

Correction: Using Conversational AI to Facilitate Mental Health Assessments and Improve Clinical Efficiency Within Psychotherapy Services: Real-World Observational Study ([e57869](#))
 Max Rollwage, Johanna Habicht, Keno Juechems, Ben Carrington, Sruthi Viswanathan, Mona Stylianou, Tobias Hauser, Ross Harper. 139

Correction: Feasibility of Multimodal Artificial Intelligence Using GPT-4 Vision for the Classification of Middle Ear Disease: Qualitative Study and Validation ([e62990](#))

Masao Noda, Hidekane Yoshimura, Takuya Okubo, Ryota Koshu, Yuki Uchiyama, Akihiro Nomura, Makoto Ito, Yutaka Takumi. 141

Tutorial

Generating Synthetic Electronic Health Record Data Using Generative Adversarial Networks: Tutorial

Chao Yan¹, PhD; Ziqi Zhang², PhD; Steve Nyemba¹, MS; Zhuohang Li², MS

¹Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN, United States

²Department of Computer Science, Vanderbilt University, Nashville, TN, United States

Corresponding Author:

Chao Yan, PhD

Department of Biomedical Informatics

Vanderbilt University Medical Center

Suite 1475, 2525 West End Ave

Nashville, TN, 37203

United States

Phone: 1 6155126877

Email: chao.yan.1@vumc.org

Abstract

Synthetic electronic health record (EHR) data generation has been increasingly recognized as an important solution to expand the accessibility and maximize the value of private health data on a large scale. Recent advances in machine learning have facilitated more accurate modeling for complex and high-dimensional data, thereby greatly enhancing the data quality of synthetic EHR data. Among various approaches, generative adversarial networks (GANs) have become the main technical path in the literature due to their ability to capture the statistical characteristics of real data. However, there is a scarcity of detailed guidance within the domain regarding the development procedures of synthetic EHR data. The objective of this tutorial is to present a transparent and reproducible process for generating structured synthetic EHR data using a publicly accessible EHR data set as an example. We cover the topics of GAN architecture, EHR data types and representation, data preprocessing, GAN training, synthetic data generation and postprocessing, and data quality evaluation. We conclude this tutorial by discussing multiple important issues and future opportunities in this domain. The source code of the entire process has been made publicly available.

(*JMIR AI* 2024;3:e52615) doi:[10.2196/52615](https://doi.org/10.2196/52615)

KEYWORDS

synthetic data generation; electronic health record; generative neural networks; tutorial

Introduction

Generating synthetic versions of private human-generated data sets has garnered increasing attention in both academia and industry as a means to enable broad data access on a large scale [1,2]. When appropriately generated, synthetic data can mirror the statistical structures of the real data upon which they are based while severing connections to real human individuals [3]. Synthetic data not only enable data sharing with minimal privacy risks but also support data augmentation (ie, artificially increase the amount of available data by generating new data) to boost the performance of machine learning (ML) models. Such a nature has significant implications for maximizing the value of patient data to improve biomedicine and health care.

The widespread adoption of electronic health record (EHR) systems has amassed vast patient data globally. Despite their potential to enrich health knowledge and support care

optimization [4-7], data accessibility remains limited due to privacy concerns [8,9], which impedes the advancement of knowledge discovery and translational artificial intelligence (AI) or ML research in health care. Synthetic data generation emerges as a solution by producing EHRs that are of minimal privacy risks while maintaining usability to facilitate endeavors [10,11] ranging from health information system (or software) testing and medical education to hypothesis generation and medical AI development. Acknowledging their benefits, multiple initiatives have relied upon synthetic data to expand the accessibility of their data for public use, including the National Institute of Health's National COVID Cohort Collaborative [12] and the Clinical Practice Research Datalink by the United Kingdom's National Institute for Health and Care Research [13].

Due in part to the limited accessibility of real EHRs, the data sets made available for biomedical research often exhibit small

sizes, insufficient diversity, missing modalities, biased subpopulation representativeness, imbalanced labels, and scarce annotations [14]. As a result, ML models trained on these data may demonstrate inferior performance, limited generalizability, and unfair outcomes (ie, when there exist disparities in model performance across patient subpopulations) [15]. Compared with solely using existing data, integrating synthetic EHR data with real data can potentially enhance model performance and reduce biases [3,16,17]. This strategy effectively enlarges the proportion of underrepresented classes or patient subpopulations within the real data and, thus, prevents the model training process from overly focusing on the dominant groups. Importantly, synthetic EHR data can be produced quickly, of arbitrary size, and at low cost, and they are able to introduce higher diversity than traditional augmentation strategies (eg, over- or undersampling), which reduces the likelihood of overfitting. It is notable that creating synthetic EHR data, when based on a private real data set and supplied to support ML innovations by a third party, offers a unique opportunity to realize the dual benefits of data sharing that maintains privacy and data augmentation.

Among numerous synthetic data generation techniques, generative adversarial networks (GANs) and their variants have showcased their capability to accurately capture the statistical properties of real EHR data while inducing low privacy risks [18-20]. GAN-based methods avoid explicitly modeling clinical knowledge and making assumptions about variables and their correlations; instead, they directly learn the underlying relationships from the multidimensional data and then generate synthetic records based on the learned model [21].

Despite the rapid advancement and evolution of synthetic EHR data generation technologies, the whole procedure for producing synthetic EHR data has not been revealed in a detailed manner. This tutorial paper aims to fill that gap by providing a sequence of step-by-step instructions, supported by complementary demo code, to assist those practitioners who are not specialized in this area to effectively translate state-of-the-art research in synthetic EHR data to practical applications. This tutorial is designed with the expectation that readers have a basic understanding of ML concepts and proficiency in Python programming. We cover multiple topics, including GAN architecture, EHR data types and matrix representation, data preprocessing, GAN training, synthetic data generation, and evaluation. For demonstration purposes, we use the state-of-the-art open-source model (ie, EMR-WGAN [22]) and a publicly available EHR data set (ie, the Medical Information Mart for Intensive Care, the Fourth Version [MIMIC-IV] [23]) for structured EHR data generation. We defer the comparisons of various GAN-based models to our

previous paper [21]. We also provide a detailed Jupyter notebook [24] to ensure the replicability of the tutorial content.

Methods

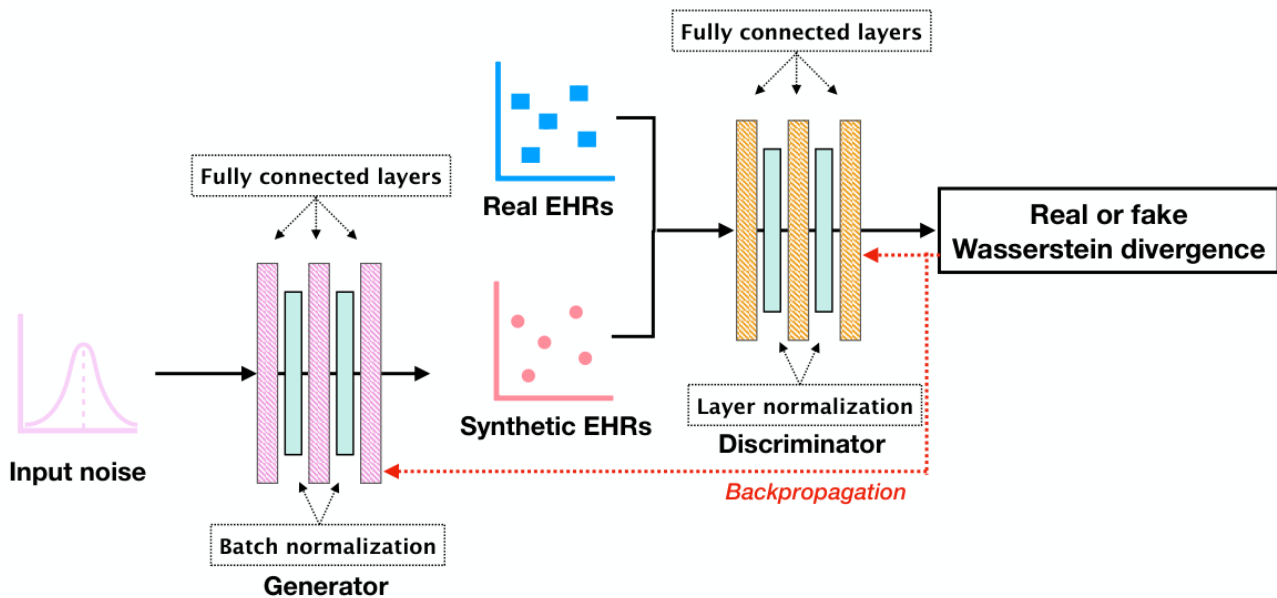
Data Set

We use the MIMIC-IV [23] data set as an example to demonstrate the generation and evaluation process of synthetic structured EHR data. MIMIC-IV is the latest version of the MIMIC EHR data, a publicly available database sourced from real EHRs of the Beth Israel Deaconess Medical Center. Adult patients admitted to the emergency department or an intensive care unit between 2008 and 2019 were incorporated. MIMIC-IV includes a wide array of information such as diagnoses, procedures, treatments, measurements, orders, free-text clinical notes, and mortality labels that indicate whether a patient died within 1 year following their last hospital stay within the timeframe. In this tutorial, we extracted patients from MIMIC-IV who had at least 1 hospital admission and were discharged alive following their last hospitalization. To build a simple demonstration data set, we extracted patients' demographic information (including age, sex, and race); diagnoses; and 2 types of the latest measurements, that is, BMI and blood pressure (systolic and diastolic pressures). We reduced the dimensionality by converting the *International Classification of Disease, Ninth or Tenth Revision (ICD-9/10)* diagnosis codes to phenome-wide association study codes (ie, phecodes), which aggregate billing codes into clinically meaningful phenotypes [25].

GAN Architecture

GANs consist of 2 neural networks: a generator that is trained to produce realistic synthetic data from random noise and a discriminator that aims to distinguish between real and synthetic data generated by the generator [26]. During the iterative training process, the generator receives feedback through backpropagation from the discriminator and then continues to refine its capability until the discriminator cannot differentiate between real and synthetic data. GAN variants retain this common architecture while customizing how each component is implemented to adapt to various data types and stabilize the training procedure [27]. Specifically, EMR-WGAN [22] (Figure 1) applies Wasserstein divergence [28] to characterize the distance between real and synthetic data and uses fully connected layers, as well as normalization techniques, to construct the generator and discriminator. This combination of design has demonstrated its superiority in capturing the statistical characteristics of real data over other models for EHR data generation [21].

Figure 1. An architectural overview of EMR-WGAN. EHR: electronic health record.



EHR Data Types and Matrix Representation

Structured EHR data for secondary analysis are usually stored in a relational database (eg, Epic Clarity) or in multiple separated files with a tabular format (eg, MIMIC-IV), where each row represents a patient's fact, such as demographic information, or a medical event marked by a timestamp, such as disease diagnoses, medication prescriptions, measurements, medical procedures, and clinical outcomes related to an encounter. These data are usually represented by continuous, categorical, or discrete variables (Figure 2A). Continuous variables can assume any value within a specific range, making them suitable for representing medical measurement results, such as hemoglobin A_{1c} readings. Discrete variables are characterized by a countable number of numerical values, such as the number of pregnancies. However, the discrete variables with a broad range of values, such as age, can be approximated as continuous variables. In contrast, categorical variables are defined by a limited and typically unchanging set of options, such as sex, race, and diagnosis. Unlike discrete variables that naturally possess an order, categorical variables typically do not have a hierarchical relationship with nonquantitative distinctions, such as classifications of "low," "medium," or "high." In the practice of synthetic data generation, discrete variables with a limited

range of values are sometimes considered categorical for simplicity.

Timestamps indicate medical events' positions on the time dimension. In the longitudinal synthetic EHR generation scenario, the time interval between 2 consecutive medical events is often used as a substitute for timestamps [29,30]. In this paper, we focus on demonstrating the generation of snapshot (or static) EHR data by removing or transforming the occurrence time of medical events so that all information about 1 patient can be represented by 1 single row of a table. While temporal information on medical events adds significant value to EHR data, snapshot EHR data still offers a wealth of information to support care delivery, data analytics, research, policy making, and education. Figure 2B shows a transformed snapshot EHR data matrix (EHR matrix for short) derived from Figure 2A. In this matrix, each row denotes a patient's record, and each column denotes a variable. It is notable that each categorical variable with k ($k > 2$) distinct options is represented by k new variables (or columns) in a one-hot manner (eg, insurance and number of pregnancies in the example), whereas the categorical variables with only 2 options (eg, mortality in the example) are represented by a single binary column.

Figure 3 illustrates the whole process of producing synthetic EHR data by training generative models.

Figure 2. An illustration of (A) data types in electronic health record data, and (B) transformed snapshot electronic health record matrix for synthetic data generation. #P: number of pregnancies; BP-D: diastolic blood pressure; BP-S: systolic blood pressure; H-A1C: hemoglobin A1C; HT: hypertension; Ins: insurance; T2D: type 2 diabetes.

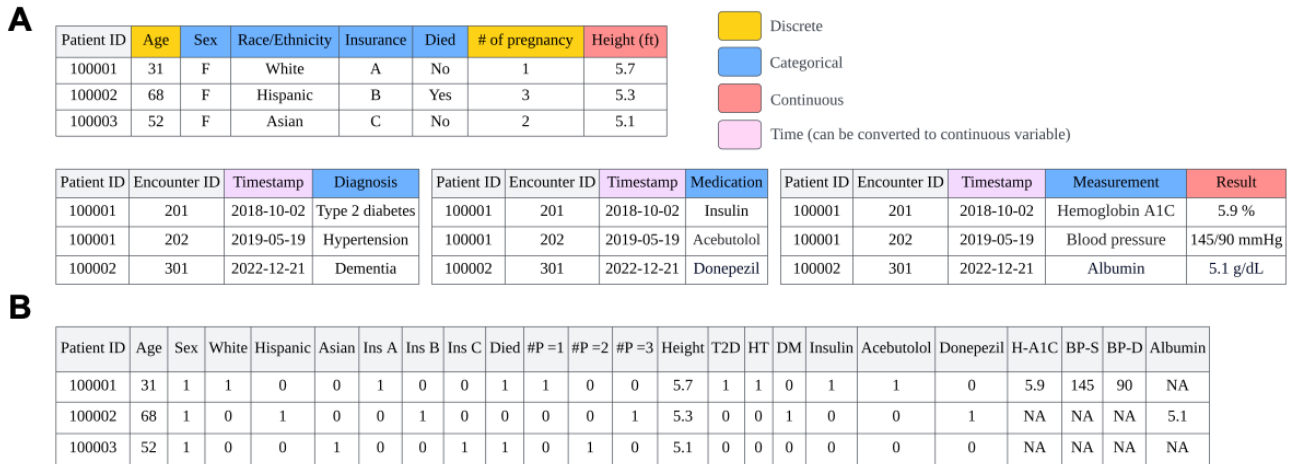
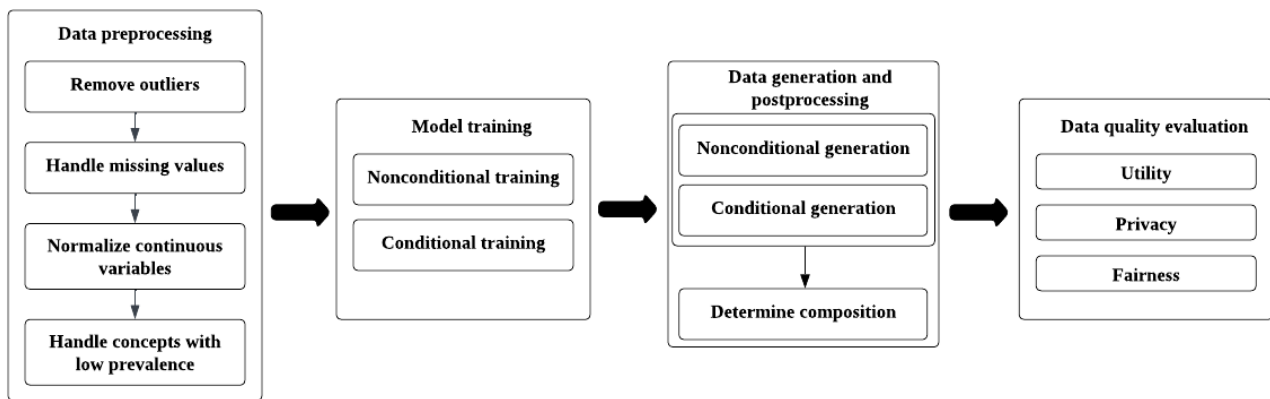


Figure 3. An overview of synthetic electronic health record data generation process through training generative models.



Data Preprocessing

Overview

With the patient cohort of interest extracted and the corresponding matrix representation of their EHR data (ie, EHR matrix) obtained, a series of data preprocessing procedures need to be performed in order to produce a GAN-ready training data set. The procedures include (1) removing outliers, (2) handling missing values, (3) normalizing continuous variables, and (4) handling concepts with low prevalence.

Removing Outliers

We define outliers in structured EHR data as data points that are significantly distant from the majority of values. These can be data points that conflict with common sense or established clinical knowledge. This phenomenon typically occurs when incorrect values are entered or generated in EHRs and is particularly prevalent among discrete and continuous variables. Outliers can also represent occurrences that are theoretically possible but exceedingly rare, which creators of synthetic data may opt to exclude depending on the requirement of data generation. In both cases, it is critical to inspect the distribution of each noncategorical variable by creating histograms and reviewing basic statistical measures, such as the mean, median, minimum, and maximum values. As an example, we examined

the distribution of BMIs in the processed EHR matrix, which led to findings that the minimum and maximum BMIs are 0 and 107,840.2. There are 366 patients with their latest BMIs greater than 60, and there are 120 patients with their BMIs less than 10. Given that these BMIs are unreasonable for adult patients, we removed the corresponding patients from the EHR matrix. One alternative solution that preserves the amount of data available for training generative models is to clip outlier values based on a pre-established reasonable range for the relevant variables.

Handling Missing Values

Multiple reasons can contribute to EHR data missingness, including, but not limited to, fragmented EHRs, incomplete documentation, data entry errors, and skipped clinical measurements. These reasons have also been classified in the literature as missing completely at random, missing at random, or missing not at random [31]. Before proceeding with imputation, it is generally recommended to eliminate variables with a high missing rate (eg, more than 50%). Numerous missing data imputation methods for EHR data have been developed [32-35], such as random sampling, prediction-based methods, and nearest neighbor-based methods. Yet, growing evidence has suggested that different methods are suitable for different missingness types, data sets, and use cases and that there is no single method that is universally considered the best for all

scenarios. In this tutorial, we applied a random sampling strategy to impute missing values in BMI, which had a 38.6% missing rate, and both diastolic and systolic blood pressure, each with a 43.5% missing rate. Specifically, we randomly sampled and then imputed values based on the marginal distribution of each variable, though we acknowledge that this might not be the optimal strategy for all use cases of this data set.

Normalizing Continuous Variables

Continuous variables each possess a specific range of values, as illustrated by the difference between blood pressure and height in feet in [Figure 2B](#). Normalizing continuous variables prevents the training of generative models from being dominated by variables with large ranges. To keep the distribution of each continuous variable, it is recommended to linearly compress their values into the range of (0,1), with its maximum and minimum values the same as binary variables. Given a continuous variable v , as well as its maximum value v_{\max} and minimum value v_{\min} , the normalized value v'_k of v_k can be calculated as:



(1)

Handling Concepts With Low Prevalence

Concepts with low prevalence correspond to clinical variables that represent rare facts or events within the patient cohort. Examples include diseases, procedures, and medications that are uncommonly diagnosed, executed, and prescribed, respectively. ML-based generative models, including GANs, cannot accurately capture the statistical properties of these variables, as well as their correlations with other variables, due to the limited observations in the real data set. Noise, however, could be induced by keeping these variables in the EHR matrix for GAN training. To address this issue, several strategies can be used as follows: (1) removing these low-prevalence variables from the EHR matrix and reintroducing them in the postprocessing stage when needed, (2) rolling up variable granularity to a higher level to raise prevalence (eg, converting raw *ICD-9/10* codes to their integer level or to phecodes), and (3) combining both approaches. In this tutorial, we converted *ICD-9/10* diagnosis codes to phecodes and then removed the phecodes with a prevalence of less than 5×10^{-5} .

Model Training

Depending on model architectures, distance measures, and training techniques used (such as batch sizes, and alternating strategies for training the generator and discriminator), GAN-based synthetic EHR data generation models show varied capabilities in capturing the properties of real data. However, they typically encounter 2 main types of uncertainties throughout the training process. First, GAN training usually occurs within a parameter space that is both complex and high-dimensional. This inherent complexity and the adversarial dynamics of GANs often lead to an unstable training process that converges to suboptimal solutions. Such nature of GAN training can cause multiple undesired phenomena, including mode collapse (the generator maps different inputs to the same output) and mode

drop (the generator only captures part of the distribution in the real data) [22]. Second, the model checkpoint that corresponds to the highest quality of the synthetic data is not necessarily the one with the lowest training loss. In addition, it has been realized that overtraining GAN-based models might degrade the quality of synthetic data. In other words, there is no monotonic relationship between training loss and the quality of synthetic data.

In order to attain the synthetic EHR data of the highest possible data quality that a GAN-based model can achieve, we highly recommend training the model multiple times (or multiple runs) from scratch and testing data quality at multiple checkpoints along the training trajectory of each run. This mechanism will not only improve the quality of synthetic EHR data to better support downstream uses but also contribute to more fair comparison between different generative models. This is crucial because researchers often need to select the best synthetic EHR generation model tailored to the real data sets and designated use cases [21].

Two different training paradigms can be considered for scenarios involving patient labels, for example, health outcomes (eg, mortality, readmission, and discharge), medical events of interest (eg, the presence of phenotypes and interventions), and patients' demographic information (eg, race, sex, and age groups). The nonconditional training paradigm does not distinguish the label variables in the EHR matrix from the remaining variables, whereas the conditional training paradigm uses the label variables to guide model training, as well as the generation of the synthetic EHR data [22], which enables the control over the categories of the generated data in terms of the label variables. Conditional training is usually achieved by incorporating the label variables as extra input of the neural networks of the generator and discriminator. However, consensus has not been established regarding which paradigm achieves a higher quality of synthetic EHR data.

When categorical variables with k ($k > 2$) unique options are converted into k binary variables within the EHR matrix, it is essential to maintain the one-hot constraint in the synthetic data. This means that only 1 of the binary variables can take a value of 1, while the remaining $k-1$ variables must be set to 0. However, the GAN training mechanism may lead to a violation of this constraint. To solve this issue, a SoftMax layer should be attached to the output of the generator to preserve the one-hot constraint.

Additionally, real data may contain critical record-level constraints that represent established clinical knowledge, which need to be preserved in the synthetic data. For instance, female patients should not be assigned male-specific diseases, such as prostate cancer. Such constraints can be effectively enforced by adding corresponding penalty terms to the loss function of GANs [36].

In this tutorial, for illustrative purposes, we use the nonconditional paradigm, preserve the one-hot constraints, yet refrain from imposing record-level constraints during model training to showcase the phenomenon of clinical knowledge violation in results.

Synthetic Data Generation and Postprocessing

Random noises, typically drawn from the standard normal distribution, need to be input into the trained generator to produce synthetic EHR data. By repeating this process, the generator is able to produce a specified quantity of synthetic records. When the conditional training paradigm is adopted, the prespecified label values should also be fed into the generator as part of the input. The capability to generate synthetic data in any desired quantity and to control the categories of the generated records affords us the flexibility to determine the composition of the resultant data set for downstream use. This nature has significant implications for data augmentation as it enables practitioners to augment their existing data sets with synthetic records tailored to their specific needs.

By applying a sigmoid or SoftMax function as the output layer of the generator, variables in the synthetic data assume values ranging between 0 and 1. For noncontinuous variables, rounding the values is necessary, whereas the values of continuous variables require rescaling to their original range by applying the inverse version of Equation 1. This process ensures that the synthetic data preserves the value ranges found in the real data set.

Data Quality Evaluation

Overview

The quality evaluation of synthetic EHR data primarily revolves around 3 key aspects: data utility, privacy, and fairness. This process requires a comparison between synthetic data and real data using a set of metrics. In this tutorial, we select multiple commonly used metrics that are complementary to each other to demonstrate data evaluation. Below, we provide a brief overview of these metrics. For more comprehensive details, we point readers to several recent publications in the field [18,19,21], which provide in-depth explanations of how these metrics are designed.

Data utility measures the usefulness and applicability of a data set for specific purposes. More concretely, it is evaluated by determining how well the generated data captures the critical characteristics present in the real EHR data. Unlike imaging data whose quality can be visually evaluated by humans or assessed using a single metric, the quality of synthetic EHR data is less intuitive and can vary in a variety of aspects. Typically, data utility is assessed by evaluating the extent to which synthetic EHR data (1) resemble the statistical characteristics of real data at both variable and record (or patient) levels and (2) retain the capability of developing ML models that perform comparably to those trained using real data. In earlier research, the concept of resemblance was often characterized as being distinct and independent from data utility. Variable-level characteristics include but are not limited to, variables' marginal distributions, their correlations, and joint distributions, whereas record-level characteristics cover multiple crucial aspects, including the violation rate of clinical knowledge, the distribution of medical concept quantity, etc.

Dimension-Wise Distribution

This metric evaluates the degree to which a synthetic data set captures the marginal distributions of variables in the real data. It calculates the average of the absolute prevalence differences (APDs) for categorical variables and the average of the Wasserstein distances for continuous variables between real and synthetic data sets. When both types of variables are present, we add these 2 values together and then normalize the sum to derive the final score, which is referred to as dimension-wise distance (DWD). A lower value of this metric indicates a higher level of data utility.

Column-Wise Correlation

This metric measures how well a synthetic data set maintains the correlations of variables present in the real data. It calculates the Pearson correlation coefficient matrices (for all variable pairs) in both the real and synthetic data sets and then computes the average of the absolute differences between corresponding cells in these 2 matrices. A lower value of this metric indicates a higher level of data utility.

Latent Cluster Analysis

This metric evaluates the effectiveness of a synthetic data set in preserving the underlying structures (or joint distribution) of real data in the latent space. It involves combining the real and synthetic EHR matrices and then applying principal component analysis to project the combined data set into a latent space that covers a specific threshold of variance in the system. Subsequently, a clustering algorithm, such as *k*-means, is used to derive the latent deviation, which is calculated as the logarithmic average of the transformed ratio of real data points present in each identified cluster. A lower value of this metric suggests a closer resemblance of the synthetic data set's latent distribution to that of the real data.

Medical Concept Abundance

This metric quantifies the degree to which a synthetic data set maintains the quantity of the record-level information in the real data. The normalized Manhattan distance between the histograms of the number of distinct record-level medical concepts for real and synthetic data sets is calculated as the medical concept abundance distance. A lower value of this metric indicates a higher level of real-synthetic data similarity.

Clinical Knowledge Violation

This metric measures the degree to which a synthetic EHR data set violates clinical knowledge, particularly in terms of maintaining record-level consistency with established medical common sense. To do so, we identified the most prevalent diagnoses (3 in this tutorial) that are only associated with 1 sex in the real data and subsequently computed the average ratio of all diagnoses appearing in the opposite sex in the synthetic data sets. A lower value of this metric indicates a higher level of data utility.

Prediction Performance

This metric evaluates the capability of a synthetic EHR data set to support ML model development. The real data set is split into a training set and a testing set. The reference model is then

trained using the real training set and evaluated on the real testing set by calculating the area under the receiver operating characteristic curve (AUROC). Subsequently, a new model is trained using the synthetic data set and then evaluated on the same real testing set. These 2 scenarios are referred to as training on real testing on real (TRTR) and training on synthetic testing on real (TSTR), respectively. The more closely the AUROC of TSTR aligns with that of TRTR, the higher the utility of the synthetic data set.

Feature Importance

This metric focuses on assessing how reliably a synthetic data set reveals key features that are significant in the prediction task. We first identified the top N (20 in this tutorial) important features in the TRTR scenario by computing the Shapley additive explanations values of all features and then computed the overlap proportion of the top N features with those identified in the TSTR scenario. The higher the proportion, the higher the data utility. Note that “feature” used in the context of feature importance is equivalent to variable.

Data privacy evaluation is crucial when considering the sharing of synthetic EHR data. While synthetic EHR data are designed to minimize privacy risks by severing the linkage to real patients, it is still important to conduct thorough privacy evaluations to ensure the preservation of individual privacy in multiple privacy inference settings, where adversaries’ knowledge and objectives differ. Across different privacy inference settings, it is commonly assumed that adversaries only have access to the generated synthetic data, but not the synthetic data generation model. Examples of widely used privacy metrics include membership inference risk and attribute inference risk [21,22,37], each with values ranging from 0 to 1. Membership inference risk measures the ability of an adversary to infer whether a specific real record is part of the data set to train the synthetic data generation model. It is quantified using the F_1 -score of the inference based on the distances between targeted records and all synthetic records. By contrast, attribute inference risk reflects an adversary’s capability to infer sensitive attributes of partially observed real EHRs. Specifically, it is calculated through the weighted sum of F_1 -scores of the inferences against sensitive attributes.

Multiple additional metrics have been created to assess privacy risks in various contexts, including meaningful identity disclosure risk [38] and nearest neighbor adversarial accuracy risk [39]. Meaningful identity disclosure risk extends the concept of identity disclosure from the context of releasing real data to the scenario of sharing synthetic data. It encompasses a comprehensive privacy risk that involves two main aspects: (1) inferring the identifiability of patients and (2) acquiring new knowledge about targeted patients. In contrast, nearest neighbor adversarial accuracy risk assesses the extent to which a synthetic data set overfits the real training data set. Specifically, it measures the difference between (1) the aggregated distance between synthetic records and those in the real testing data set and (2) the aggregated distance between synthetic records and those in the real training data set.

Synthetic EHR data are also anticipated to fairly represent patient subpopulations with respect to protected attributes, such as age groups, sex, race, and ethnicity. Distributional differences or distances between real and synthetic data with respect to the protected attributes of interest are often used as metrics to evaluate fair representation [40]. To ensure fair data quality, synthetic data may need to show similar variations in preserving data utility and protecting privacy for each patient subpopulation, akin to their real data counterparts. This consideration of fairness requires that utility and privacy evaluations of synthetic data should be performed independently within each subpopulation and then compared across them. Another fairness consideration necessitates that synthetic data sets provide equal support for downstream AI or ML tasks across all subpopulations, regardless of the basis of the real data. Due to the complexity surrounding fairness and the absence of clear guidelines for evaluating it in synthetic EHR data, we will skip this evaluation in our demonstration.

It is crucial to note that quality evaluation of synthetic EHR data should be tailored to align with specific use cases because different use cases prioritize the preservation of different data aspects. For instance, when the synthetic EHR data are intended to facilitate hypothesis generation to support medical research in a controlled research environment, the evaluation would emphasize metrics that measure disease prevalence and correlations between features and outcomes, while privacy risks may be of lesser concern. On the other hand, if the synthetic EHR data are developed to support the development of clinical decision support software by third-party developers, evaluating privacy risks becomes more critical than determining whether the synthetic data preserves the nuanced statistical properties of the real data. Our previous research provides a use case-oriented benchmarking framework to enable systematic comparisons of synthetic data generation models [21]. The users of this framework determine the prioritization of evaluation metrics by providing a weight profile, which applies to the evaluation results from individual metrics and represents the relative importance or preference assigned to each metric. The final score of a synthetic data set or a synthetic data generation model is derived by aggregating the weighted results for all considered metrics.

Using this benchmarking framework enables the selection of the most suitable synthetic data set for a specific use case or the comparison of various synthetic data generation models (not necessarily limited to those that are GAN-based) based on the scores assigned to produced synthetic data sets.

Results

Overview

In this section, we present the results of data quality evaluation for synthetic EHR data sets in terms of data utility and privacy. Furthermore, we demonstrate how to compare these synthetic EHR data sets to identify the most suitable one for specific use cases. To do so, 70% of records of the preprocessed MIMIC-IV data set were used to train the EMR-WGAN model and the remaining 30% of records were used for evaluation purposes. Considering the inherent uncertainties associated with

GAN-based model training as mentioned earlier, EMR-WGAN was independently trained 5 times. While we recommend examining multiple checkpoints during each model's training phase, for the purposes of this demonstration, we selected an epoch with a relatively low training loss from each independent training session to generate the corresponding synthetic data set. All synthetic data sets produced by these models have the same size as the real training data set. The complete process of data quality evaluation can be found in the shared Jupyter notebook [24].

Characteristics of the Real Data Set

Table 1 provides an overview of the basic characteristics of the MIMIC-IV cohort selected for the creation and evaluation of

synthetic EHR data. We initially included a total of 181,294 patients who had at least 1 hospital admission and were discharged alive for their last hospital stays. The average age of this cohort is 56.2 (SD 20.4) years. This cohort comprises 96,617 (53.3%) female individuals and multiple racial groups, with 7667 (4.2%) Asian; 23,999 (13.2%) Black; 10,058 (5.5%) Hispanic; 121,954 (67.3%) White; 10,078 (5.6%) belonging to other races; and 7538 (4.2%) of unknown race. A total of 20,493 (11.3%) of the cohort died within 1 year after their last hospital stay. The data preprocessing procedure led to the removal of 548 patients and more reasonable distributions of BMI, diastolic, and systolic blood pressures. The curated real EHR matrix contains 1460 columns after we removed 140 extremely rare diagnoses.

Table 1. Cohort characteristics before and after data preprocessing.

Characteristics	Distributions and values	
	Before preprocessing (n=181,294)	After preprocessing (n=180,746)
Cohort size, n (%)	181,294 (100)	180,746 (100)
Age (y), mean (SD)	56.2 (20.4)	56.2 (20.3)
Sex, n (%)		
Female	96,617 (53.3)	96,304 (53.3)
Male	84,677 (46.7)	84,442 (46.7)
Race, n (%)		
Asian	7667 (4.2)	7654 (4.2)
Black	23,999 (13.2)	23,889 (13.2)
Hispanic	10,058 (5.5)	10,035 (5.6)
White	121,954 (67.3)	121,603 (67.3)
Others	10,078 (5.6)	10,049 (5.6)
Unknown	7538 (4.2)	7516 (4.2)
Died within 1 year, n (%)	20,493 (11.3)	20,414 (11.3)
BMI, mean (SD)	21.1 (27.03)	28.4 (6.8)
Diastolic blood pressure, mean (SD)	47.6 (36.4)	73.6 (11.8)
Systolic blood pressure, mean (SD)	81.9 (62.3)	126.6 (18.2)
Top 10 prevalent diagnoses (in phecodes), n (%)		
Hypertension (401)	57,238 (31.6)	57,056 (31.6)
Disorders of lipid metabolism (272)	39,216 (21.6)	39,103 (21.6)
Other anemias (285)	33,979 (18.7)	33,844 (18.7)
Essential hypertension (401.1)	31,694 (17.5)	31,541 (17.5)
Hyperlipidemia (272.1)	28,011 (15.5)	27,896 (15.4)
Diseases of esophagus (530)	25,887 (14.3)	25,800 (14.3)
Cardiac dysrhythmias (427)	25,284 (14)	25,195 (13.9)
Mood disorders (296)	25,201 (13.9)	25,089 (13.9)
Tobacco use disorder (318)	24,152 (13.3)	24,054 (13.3)
Disorders of fluid, electrolyte, and acid-base balance (276)	23,895 (13.2)	23,807 (13.2)
Diabetes mellitus (250)	23,789 (13.1)	23,695 (13.1)
Total number of columns in electronic health record matrix	1600	1460

Data Utility

Figure 4 illustrates the dimension-wise distribution results and the associated APD for categorical variables. Although all 5 runs effectively maintain the marginal distributions of these variables, the second run exhibits the smallest APD. When considering both the categorical and continuous variables (ie, age, BMI, diastolic, and systolic blood pressures), the second run still achieves the lowest DWD. By contrast, the third run is associated with the highest DWD, indicating a relatively low effectiveness in preserving dimension-wise distributions.

Figure 5 summarizes the evaluation results of the 5 synthetic data sets for the remaining 6 data utility metrics, with the indication of directional implications of the values under each

metric. Notably, the second run demonstrates the highest data utility in column-wise correlation, latent cluster analysis, prediction performance, and feature importance and secures the second position in medical concept abundance. Yet, its score in clinical knowledge violation is positioned fourth. Additionally, it was observed that male-specific diagnoses are more than 10 times as likely to be incorrectly assigned to female records in the synthetic data sets compared with similar violations for female-specific diagnoses. This suggests that the correlations between sex and sex-specific diagnosis columns were not equally preserved, possibly resulting from different levels of complexity (or noise) in the data pertaining to different sexes. While this phenomenon falls beyond the scope of this tutorial, it merits further exploration.

Figure 4. Dimension-wise distribution for categorical variables. The dashed diagonal line indicates the perfect replication of variable prevalence. APD: absolute prevalence difference; DWD: dimension-wise distance.

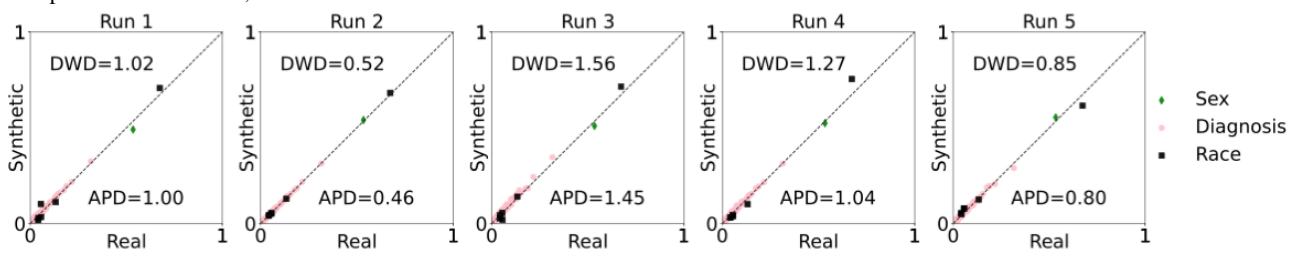
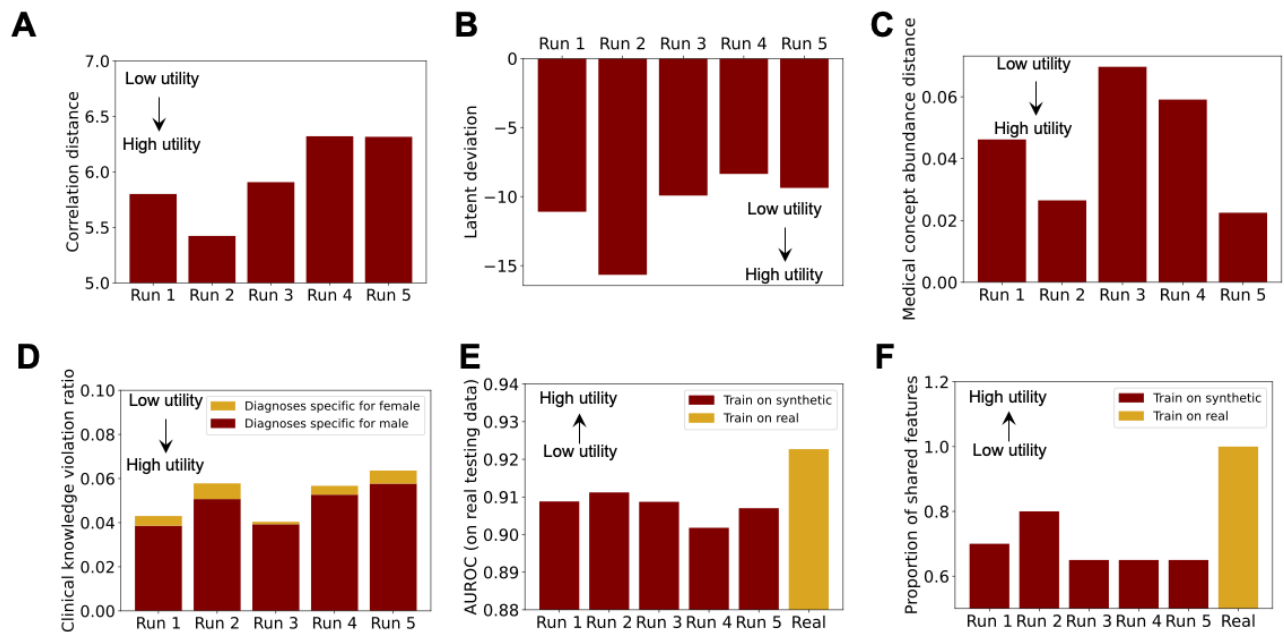


Figure 5. Data utility in (A) column-wise correlation, (B) latent cluster analysis, (C) medical concept abundance, (D) clinical knowledge violation, (E) prediction performance, and (F) feature importance. For clinical knowledge violation, “hyperplasia of prostate,” “cancer of prostate,” and “erectile dysfunction” are examined as male-specific diagnoses (in phecodes); “other conditions or status of the mother complicating pregnancy, childbirth, or the puerperium,” “known or suspected fetal abnormality affecting management of mother,” and “other complications of pregnancy necrotizing enterocolitis” are examined as female-specific diagnoses (in phecodes). AUROC: area under the receiver operating characteristic curve.



Privacy

Table 2 presents the privacy risk associated with each synthetic EHR data set in terms of membership inference attack and attribute inference attack. It also includes a baseline comparison, which corresponds to an extreme situation of releasing real data.

Compared with the real data set, every synthetic data set achieves substantially reduced risks. While the variance in risk levels among the 5 synthetic data sets is relatively small, the second run exhibits the highest membership inference risk and the second lowest risk in attribute inference.

Table 2. Privacy risks of synthetic electronic health record data sets. For each risk category, the identical risk value is attributed to a loss of precision.

Risk type	Run 1	Run 2	Run 3	Run 4	Run 5	Real
Membership inference	0.29	0.31	0.29	0.29	0.30	0.91
Attribute inference	0.14	0.14	0.14	0.13	0.14	0.97

Identifying the Most Suitable Synthetic Data Set for a Specific Use Case

We have obtained the evaluation results of all 5 synthetic data sets for individual metrics, allowing for straightforward derivation of their rankings in each metric as presented in Table 3. A smaller ranking position indicates better data quality. In this tutorial, we consider two distinct use cases of synthetic EHR data: (1) ML model development, which prioritizes the performance of prediction tasks and model explainability, and

(2) education, which focuses more on the record-level consistency with clinical knowledge, prevalence of diagnoses, and privacy. We proposed example weight profiles for these 2 use cases and then calculated the overall rankings of the synthetic data sets for each scenario. The analysis identifies the second and third runs as the most suitable data sets for ML development and education, respectively. This observation further justifies that the quality evaluation of synthetic data should be in the context of use cases.

Table 3. Data quality rankings of synthetic data sets. Weight profiles A and B correspond to the use cases for supporting machine learning model development and education, respectively. Overall rankings of data sets are weighted summation of individual rankings in all metrics.

Metric	Weight profile A	Weight profile B	Run 1	Run 2	Run 3	Run 4	Run 5
Utility							
Dimension-wise distribution	0.1	0.1	3	1	5	4	2
Column-wise correlation	0.1	0.1	2	1	3	5	4
Latent cluster analysis	0.1	0.0	2	1	3	5	4
Medical concept abundance	0.0	0.0	3	2	5	4	1
Clinical knowledge violation	0.1	0.4	2	4	1	3	5
Prediction performance	0.2	0.0	2	1	3	5	4
Feature importance	0.2	0.0	2	1	4	4	4
Privacy							
Membership inference	0.1	0.2	3	5	2	1	4
Attribute inference	0.1	0.2	3	2	4	1	5
Overall rankings for weight profile A	N/A ^a	N/A	2.3	1.8 ^b	3.2	3.7	4.0
Overall rankings for weight profile B	N/A	N/A	2.5	3.2	2.4 ^b	2.5	4.4

^aN/A: not applicable.

^bIndicates the most suitable data set for each use case.

Discussion

Principal Findings

GAN-based synthetic data generation has demonstrated significant potential to enlarge the accessibility of health data and enhance the effectiveness of ML in health care [41-43]. This tutorial demonstrates how to create and evaluate structured synthetic EHR data by applying a GAN-based generative model to a publicly available EHR data set. Beyond introducing technical details, we aim to discuss several important issues related to this topic.

GAN-based synthetic EHR data generation models exhibit limited capability in accurately representing and then generating the concepts with low prevalence. This is also a common challenge for almost all ML methods. From our experience, incorporating these concepts into the real data for GAN training, compared with removing them, can result in adverse effects on

capturing the distributions of prevalent concepts. In settings where accurate representation of concepts with low prevalence is crucial (eg, synthetic data are developed to replicate studies related to rare diseases), additional efforts should be dedicated to ensuring their fidelity in the synthetic data. One solution is to increase the representation of these concepts in the real data through data collection or data oversampling. The second solution is to independently model the cohort associated with the targeted concept. Subsequently, the synthetic data for this specific cohort can be generated and then merged with the main synthetic data. Another approach, which is modeling-free, is to perturb the real EHR data with the targeted concept based on expert knowledge and then add the resultant data back into the main synthetic data. It should be noted that the quality of synthetic data after using these approaches should be comprehensively evaluated.

Selecting the most suitable synthetic EHR data set or synthetic data generation model for a targeted use case is subject to 2 types of tradeoffs: extrinsic and intrinsic tradeoffs. Users of this technology control the extrinsic tradeoff by prioritizing which aspects of the data to preserve in data quality evaluation. This can be accomplished by using an appropriate set of evaluation metrics and assigning weights to each metric to achieve a balanced evaluation outcome that aligns with the use case, as mentioned earlier. Different prioritization strategies can yield variations in evaluation results, thereby influencing the selection of the optimal data set or model.

The intrinsic tradeoff arises from the inherent interrelation and tension among data utility, privacy, and fairness. In general, better data utility aligns with a more accurate representation of the nuanced statistical characteristics present in the real data, which can, in turn, improve the success rate of privacy inference regarding sensitive information about patients. Similarly, aiming for a higher level of privacy protection is often paired with a reduction in data fidelity. Different synthetic EHR generation models, and even different runs of the same model, can exhibit varying utility-privacy tradeoffs. The choices of model structures, parameter settings, data preprocessing, and learning methods can all impact the resulting tradeoff. In addition, one can integrate privacy protection strategies during model training, such as differential privacy, to induce more privacy protection. However, for the use cases that demand high fidelity of synthetic EHR data, such as data analysis or augmenting medical AI development, the integration of additional privacy safeguards may potentially limit the value of synthetic data for the intended scenarios.

Pursuing either a higher overall utility of synthetic EHR data or stronger privacy may lead to poor fairness across patient subpopulations. This is because different patient subpopulations may not be equally affected and that the unique characteristics of underrepresented groups are more likely to be neglected. Similarly, focusing solely on fairness may result in a lower level of overall data utility or privacy. As such, both extrinsic and intrinsic tradeoffs among data utility, privacy, and fairness impact the determination of the most suitable synthetic EHR data or synthetic EHR data generation model for a specific use case.

Multiple key questions regarding the best practice of synthetic EHR data generation remain unanswered in the literature. First, the determination of the appropriate size of real data needed to train GANs and other generative models for a specific data generation task, along with an effective estimation approach, is uncertain and lacks comprehensive research. Second, the scalability of GANs and other generative models with respect to varying sizes of the variable space is still not well understood. Third, the optimal matrix representations of various EHR data types, in particular when mixed together, are relatively unexplored in current research. All of these questions need to be answered through systematic research.

The evolution of synthetic EHR data generation technology presents numerous opportunities for various applications and advancements. We conclude this paper by highlighting several

future research directions that are worth exploring and summarizing the limitations of this tutorial.

Most cutting-edge approaches for structured synthetic data generation, including EHR data, rely on a matrix or tabular representation of the real data, which involves merging all information into a single table as part of data preprocessing. When addressing the emerging need to generate a synthetic version of a relational EHR database, where patients' data are distributed in multiple tables, such as the widely adopted OMOP common data model, joining relevant tables together can lead to an unmanageable data size with significant redundancy. There is a strong need for a novel synthetic EHR data generation paradigm that can directly learn from the original database, including its structural relationships, to address the current limitations in the field.

EHR data, in a broad sense, encompass multiple modalities, including structured health information, textual notes, medical imaging data, genetic information, and more. Current synthetic EHR data generation algorithms are designed to handle a single modality at a time, leading to a lack of consistency between separately generated data when attempting to describe the same patient. Methodology innovations are required to effectively harmonize the available modalities in EHR data during model training and then generate synthetic data that cover and represent these modalities. The core objective of this task is to learn an accurate latent representation of a patient across different modalities.

Since 2023, large language models, such as OpenAI's ChatGPT and Google's Med-PaLM 2, have gained substantial attention due to their remarkable ability to generate high-quality free text responses to users' questions and instructions. Such exceptional ability stems from their extensive pretraining on vast amounts of textual data, which contain a wide range of human knowledge and common sense. In addition, the users of these models can demand the desired format of their output such as CSV and JSON. This entails a new opportunity for synthetic EHR data generation. While private EHR data have not been used by these models, an appropriate fine-tuning process using real EHR data can quickly shape them into synthetic EHR data generators. Compared with other generative methods, large language models could potentially strengthen the generation of synthetic EHR data in multiple critical aspects. First, large language models have encoded complex knowledge and relationships between medical concepts through extensive pretraining. When fine-tuned on real EHR data sets, they can more easily capture the nuances in intricate patient data and understand the underlying data semantics, which would not be easily achieved by other generative models. Second, large language models can generate data with stronger contextual relevance and coherence. In other words, they are more capable of producing data that are not only syntactically and semantically correct but also consistent with real-world scenarios and knowledge. Third, with prompt-level customization, these models can be tailored to generate specific types of EHR data in a more flexible and efficient manner, significantly reducing the human effort required in postprocessing compared with previous methods.

This tutorial has several limitations. First, it focuses on simulating static structured EHR data and neglects the timestamping of medical events. However, it is important to note that EHR data inherently consists of time series, where the temporal information is critical for numerous applications, such as modeling the progression of diseases. To address this, multiple generative models have been developed to produce temporal EHR data, a process that shares similar principles to those demonstrated in this tutorial. Second, the real data set we used for demonstration purposes does not fully capture the complexity inherent in real snapshot EHR data. It is likely that a transformed snapshot EHR matrix contains a subset of columns governed by complex semantic constraints, which may not be straightforward to implement during model training. For example, a snapshot EHR matrix for a women's health cohort may include columns indicating the age and method (nature vs cesarean) for each childbirth. This scenario compounds constraints in several aspects, including patterns of missing data (eg, the data set might not contain only a record of the second delivery), the age at each delivery (eg, ages for subsequent

deliveries should be older than previous ones), and time intervals between deliveries (eg, there should be a minimum gap of 10 months between each). Addressing this type of complex constraint is still an open research question and needs more investigation.

Conclusions

Creating synthetic EHR data has been increasingly pursued to address the limited availability of real EHR data to facilitate various endeavors in the health domain. This tutorial provides a comprehensive guide to the entire process of generating synthetic structured EHR data using GANs, ranging from data representation, preprocessing, model training, and postprocessing to data generation and evaluation. By following this tutorial, as well as the open-sourced example based on the MIMIC-IV data set, we anticipate that potential users of synthetic data generation technology can understand and implement all involved components, and then correctly evaluate the produced data sets and interpret the evaluation results to fulfill their data needs.

Conflicts of Interest

None declared.

References

1. Arora A, Arora A. Synthetic patient data in health care: a widening legal loophole. *Lancet* 2022;399(10335):1601-1602. [doi: [10.1016/S0140-6736\(22\)00232-X](https://doi.org/10.1016/S0140-6736(22)00232-X)] [Medline: [35358423](https://pubmed.ncbi.nlm.nih.gov/35358423/)]
2. Gonzales A, Guruswamy G, Smith SR. Synthetic data in health care: a narrative review. *PLOS Digit Health* 2023;2(1):e0000082 [FREE Full text] [doi: [10.1371/journal.pdig.0000082](https://doi.org/10.1371/journal.pdig.0000082)] [Medline: [36812604](https://pubmed.ncbi.nlm.nih.gov/36812604/)]
3. Chen RJ, Lu MY, Chen TY, Williamson DFK, Mahmood F. Synthetic data in machine learning for medicine and healthcare. *Nat Biomed Eng* 2021;5(6):493-497 [FREE Full text] [doi: [10.1038/s41551-021-00751-8](https://doi.org/10.1038/s41551-021-00751-8)] [Medline: [34131324](https://pubmed.ncbi.nlm.nih.gov/34131324/)]
4. Beam AL, Kohane IS. Big data and machine learning in health care. *JAMA* 2018;319(13):1317-1318. [doi: [10.1001/jama.2017.18391](https://doi.org/10.1001/jama.2017.18391)] [Medline: [29532063](https://pubmed.ncbi.nlm.nih.gov/29532063/)]
5. Mooney SJ, Pejaver V. Big data in public health: terminology, machine learning, and privacy. *Annu Rev Public Health* 2018;39(1):95-112 [FREE Full text] [doi: [10.1146/annurev-publhealth-040617-014208](https://doi.org/10.1146/annurev-publhealth-040617-014208)] [Medline: [29261408](https://pubmed.ncbi.nlm.nih.gov/29261408/)]
6. Shah ND, Steyerberg EW, Kent DM. Big data and predictive analytics: recalibrating expectations. *JAMA* 2018;320(1):27-28. [doi: [10.1001/jama.2018.5602](https://doi.org/10.1001/jama.2018.5602)] [Medline: [29813156](https://pubmed.ncbi.nlm.nih.gov/29813156/)]
7. Denny JC, Ritchie MD, Basford MA, Pulley JM, Bastarache L, Brown-Gentry K, et al. PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations. *Bioinformatics* 2010;26(9):1205-1210 [FREE Full text] [doi: [10.1093/bioinformatics/btq126](https://doi.org/10.1093/bioinformatics/btq126)] [Medline: [20335276](https://pubmed.ncbi.nlm.nih.gov/20335276/)]
8. Price WN, Cohen IG. Privacy in the age of medical big data. *Nat Med* 2019;25(1):37-43 [FREE Full text] [doi: [10.1038/s41591-018-0272-7](https://doi.org/10.1038/s41591-018-0272-7)] [Medline: [30617331](https://pubmed.ncbi.nlm.nih.gov/30617331/)]
9. Cohen IG, Mello MM. Big data, big tech, and protecting patient privacy. *JAMA* 2019;322(12):1141-1142. [doi: [10.1001/jama.2019.11365](https://doi.org/10.1001/jama.2019.11365)] [Medline: [31397838](https://pubmed.ncbi.nlm.nih.gov/31397838/)]
10. Tucker A, Wang Z, Rotalinti Y, Myles P. Generating high-fidelity synthetic patient data for assessing machine learning healthcare software. *NPJ Digit Med* 2020;3(1):147 [FREE Full text] [doi: [10.1038/s41746-020-00353-9](https://doi.org/10.1038/s41746-020-00353-9)] [Medline: [33299100](https://pubmed.ncbi.nlm.nih.gov/33299100/)]
11. James S, Harbron C, Branson J, Sundler M. Synthetic data use: exploring use cases to optimise data utility. *Discover Artif Intell* 2021;1(1):15 [FREE Full text] [doi: [10.1007/s44163-021-00016-y](https://doi.org/10.1007/s44163-021-00016-y)]
12. Haendel MA, Chute CG, Bennett TD, Eichmann DA, Guinney J, Kibbe WA, et al. The National COVID Cohort Collaborative (N3C): rationale, design, infrastructure, and deployment. *J Am Med Inform Assoc* 2021;28(3):427-443 [FREE Full text] [doi: [10.1093/jamia/ocaa196](https://doi.org/10.1093/jamia/ocaa196)] [Medline: [32805036](https://pubmed.ncbi.nlm.nih.gov/32805036/)]
13. Wang Z, Myles P, Tucker A. 2019 Presented at: 2019 IEEE 32nd International Symposium on Computer-Based Medical Systems (CBMS); June 5-7, 2019; Cordoba, Spain p. 126-131. [doi: [10.1109/cbms.2019.00036](https://doi.org/10.1109/cbms.2019.00036)]
14. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. *Nat Med* 2019;25(1):44-56. [doi: [10.1038/s41591-018-0300-7](https://doi.org/10.1038/s41591-018-0300-7)] [Medline: [30617339](https://pubmed.ncbi.nlm.nih.gov/30617339/)]
15. Norori N, Hu Q, Aellen FM, Faraci FD, Tzovara A. Addressing bias in big data and AI for health care: a call for open science. *Patterns (N Y)* 2021;2(10):100347 [FREE Full text] [doi: [10.1016/j.patter.2021.100347](https://doi.org/10.1016/j.patter.2021.100347)] [Medline: [34693373](https://pubmed.ncbi.nlm.nih.gov/34693373/)]

16. Li J, Cairns BJ, Li J, Zhu T. Generating synthetic mixed-type longitudinal electronic health records for artificial intelligent applications. *NPJ Digit Med* 2023;6(1):98 [FREE Full text] [doi: [10.1038/s41746-023-00834-7](https://doi.org/10.1038/s41746-023-00834-7)] [Medline: [37244963](https://pubmed.ncbi.nlm.nih.gov/37244963/)]
17. Cui L, Biswal S, Glass LM, Lever G, Sun J, Xiao C. CONAN: complementary pattern augmentation for rare disease detection. *Proc AAAI Conf Artif Intell* 2020;34(01):614-621. [doi: [10.1609/aaai.v34i01.5401](https://doi.org/10.1609/aaai.v34i01.5401)]
18. Hernandez M, Epelde G, Alberdi A, Cilla R, Rankin D. Synthetic data generation for tabular health records: a systematic review. *Neurocomputing* 2022;493:28-45. [doi: [10.1016/j.neucom.2022.04.053](https://doi.org/10.1016/j.neucom.2022.04.053)]
19. Ghosheh G, Li J, Zhu T. A review of generative adversarial networks for electronic health records: applications, evaluation measures and data sources. *ArXiv Preprint* posted online on December 14 2022. [doi: [10.48550/arXiv.2203.07018](https://doi.org/10.48550/arXiv.2203.07018)]
20. Choi E, Biswal S, Malin B, Duke J, Stewart WF, Sun J. Generating multi-label discrete patient records using generative adversarial networks. 2017 Presented at: Proceedings of the 2nd Machine Learning for Healthcare Conference; August 18-19, 2017; Boston, Massachusetts p. 286-305. [doi: [10.48550/arXiv.1703.06490](https://doi.org/10.48550/arXiv.1703.06490)]
21. Yan C, Yan Y, Wan Z, Zhang Z, Omberg L, Guinney J, et al. A multifaceted benchmarking of synthetic electronic health record generation models. *Nat Commun* 2022;13(1):7609 [FREE Full text] [doi: [10.1038/s41467-022-35295-1](https://doi.org/10.1038/s41467-022-35295-1)] [Medline: [36494374](https://pubmed.ncbi.nlm.nih.gov/36494374/)]
22. Zhang Z, Yan C, Mesa DA, Sun J, Malin BA. Ensuring electronic medical record simulation through better training, modeling, and evaluation. *J Am Med Inform Assoc* 2020;27(1):99-108 [FREE Full text] [doi: [10.1093/jamia/ocz161](https://doi.org/10.1093/jamia/ocz161)] [Medline: [31592533](https://pubmed.ncbi.nlm.nih.gov/31592533/)]
23. Johnson AEW, Bulgarelli L, Shen L, Gayles A, Shammout A, Horng S, et al. MIMIC-IV, a freely accessible electronic health record dataset. *Sci Data* 2023;10(1):1 [FREE Full text] [doi: [10.1038/s41597-022-01899-x](https://doi.org/10.1038/s41597-022-01899-x)] [Medline: [36596836](https://pubmed.ncbi.nlm.nih.gov/36596836/)]
24. A tutorial for generating and evaluating synthetic health data based on MIMIC-IV V2.0 dataset and EMR-WGAN. GitHub, Inc. URL: https://github.com/yanchao0222/tutorial_data_synthesis_and_evaluation [accessed 2024-03-29]
25. Wu P, Gifford A, Meng X, Li X, Campbell H, Varley T, et al. Mapping ICD-10 and ICD-10-CM codes to phecodes: workflow development and initial evaluation. *JMIR Med Inform* 2019;7(4):e14325 [FREE Full text] [doi: [10.2196/14325](https://doi.org/10.2196/14325)] [Medline: [31553307](https://pubmed.ncbi.nlm.nih.gov/31553307/)]
26. Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, et al. Generative adversarial networks. *Commun ACM* 2020;63(11):139-144. [doi: [10.1145/3422622](https://doi.org/10.1145/3422622)]
27. Pan Z, Yu W, Yi X, Khan A, Yuan F, Zheng Y. Recent progress on generative adversarial networks (GANs): a survey. *IEEE Access* 2019;7:36322-36333. [doi: [10.1109/access.2019.2905015](https://doi.org/10.1109/access.2019.2905015)]
28. Gulrajani I, Ahmed F, Arjovsky M, Dumoulin V, Courville AC. Improved training of Wasserstein GANs. 2017 Presented at: Advances in Neural Information Processing Systems 30 (NIPS 2017); December 4-9, 2017; Long Beach, California, USA p. 5767-5777. [doi: doi.org/10.48550/arXiv.1704.00028]
29. Zhang Z, Yan C, Lasko TA, Sun J, Malin BA. SynTEG: a framework for temporal structured electronic health data simulation. *J Am Med Inform Assoc* 2021;28(3):596-604 [FREE Full text] [doi: [10.1093/jamia/ocaa262](https://doi.org/10.1093/jamia/ocaa262)] [Medline: [33277896](https://pubmed.ncbi.nlm.nih.gov/33277896/)]
30. Zhang Z, Yan C, Malin BA. Keeping synthetic patients on track: feedback mechanisms to mitigate performance drift in longitudinal health data simulation. *J Am Med Inform Assoc* 2022;29(11):1890-1898 [FREE Full text] [doi: [10.1093/jamia/ocac131](https://doi.org/10.1093/jamia/ocac131)] [Medline: [35927974](https://pubmed.ncbi.nlm.nih.gov/35927974/)]
31. Haneuse S, Arterburn D, Daniels MJ. Assessing missing data assumptions in EHR-based studies: a complex and underappreciated task. *JAMA Netw Open* 2021;4(2):e210184 [FREE Full text] [doi: [10.1001/jamanetworkopen.2021.0184](https://doi.org/10.1001/jamanetworkopen.2021.0184)] [Medline: [33635321](https://pubmed.ncbi.nlm.nih.gov/33635321/)]
32. Beaulieu-Jones BK, Lavage DR, Snyder JW, Moore JH, Pendergrass SA, Bauer CR. Characterizing and managing missing structured data in electronic health records: data analysis. *JMIR Med Inform* 2018;6(1):e11 [FREE Full text] [doi: [10.2196/medinform.8960](https://doi.org/10.2196/medinform.8960)] [Medline: [29475824](https://pubmed.ncbi.nlm.nih.gov/29475824/)]
33. Luo Y. Evaluating the state of the art in missing data imputation for clinical data. *Brief Bioinform* 2022;23(1):bbab489 [FREE Full text] [doi: [10.1093/bib/bbab489](https://doi.org/10.1093/bib/bbab489)] [Medline: [34882223](https://pubmed.ncbi.nlm.nih.gov/34882223/)]
34. Zhang X, Yan C, Gao C, Malin BA, Chen Y. Predicting missing values in medical data via XGBoost regression. *J Healthc Inform Res* 2020;4(4):383-394 [FREE Full text] [doi: [10.1007/s41666-020-00077-1](https://doi.org/10.1007/s41666-020-00077-1)] [Medline: [33283143](https://pubmed.ncbi.nlm.nih.gov/33283143/)]
35. Yan C, Gao C, Zhang X, Chen Y, Malin B. Deep imputation of temporal data. 2019 Presented at: 2019 IEEE International Conference on Healthcare Informatics (ICHI); June 10-13, 2019; Xi'an, China p. 1-3. [doi: [10.1109/ichi.2019.8904776](https://doi.org/10.1109/ichi.2019.8904776)]
36. Yan C, Zhang Z, Nyemba S, Malin BA. Generating electronic health records with multiple data types and constraints. *AMIA Annu Symp Proc* 2020;2020:1335-1344 [FREE Full text] [Medline: [33936510](https://pubmed.ncbi.nlm.nih.gov/33936510/)]
37. Zhang Z, Yan C, Malin BA. Membership inference attacks against synthetic health data. *J Biomed Inform* 2022;125:103977 [FREE Full text] [doi: [10.1016/j.jbi.2021.103977](https://doi.org/10.1016/j.jbi.2021.103977)] [Medline: [34920126](https://pubmed.ncbi.nlm.nih.gov/34920126/)]
38. El Emam K, Mosquera L, Bass J. Evaluating identity disclosure risk in fully synthetic health data: model development and validation. *J Med Internet Res* 2020;22(11):e23139 [FREE Full text] [doi: [10.2196/23139](https://doi.org/10.2196/23139)] [Medline: [33196453](https://pubmed.ncbi.nlm.nih.gov/33196453/)]
39. Yale A, Dash S, Dutta R, Guyon I, Pavao A, Bennett KP. Generation and evaluation of privacy preserving synthetic health data. *Neurocomputing* 2020;416:244-255. [doi: [10.1016/j.neucom.2019.12.136](https://doi.org/10.1016/j.neucom.2019.12.136)]
40. Bhanot K, Qi M, Erickson JS, Guyon I, Bennett KP. The problem of fairness in synthetic healthcare data. *Entropy (Basel)* 2021;23(9):1165 [FREE Full text] [doi: [10.3390/e23091165](https://doi.org/10.3390/e23091165)] [Medline: [34573790](https://pubmed.ncbi.nlm.nih.gov/34573790/)]

41. Hashemi AS, Soliman A, Lundström J, Etmnani K. Domain knowledge-driven generation of synthetic healthcare data. *Stud Health Technol Inform* 2023;302:352-353. [doi: [10.3233/SHTI230136](https://doi.org/10.3233/SHTI230136)] [Medline: [37203680](https://pubmed.ncbi.nlm.nih.gov/37203680/)]
42. Armanious K, Jiang C, Fischer M, Küstner T, Hepp T, Nikolaou K, et al. MedGAN: medical image translation using GANs. *Comput Med Imaging Graph* 2020;79:101684. [doi: [10.1016/j.compmedimag.2019.101684](https://doi.org/10.1016/j.compmedimag.2019.101684)] [Medline: [31812132](https://pubmed.ncbi.nlm.nih.gov/31812132/)]
43. Hashemi AS, Etmnani K, Soliman A, Hamed O, Lundström J. Time-series anonymization of tabular health data using generative adversarial network. 2023 Presented at: 2023 International Joint Conference on Neural Networks (IJCNN); June 18-23, 2023; Gold Coast, Australia p. 1-8. [doi: [10.1109/ijcnn54540.2023.10191367](https://doi.org/10.1109/ijcnn54540.2023.10191367)]

Abbreviations

AI: artificial intelligence

APD: absolute prevalence difference

AUROC: area under the receiver operating characteristic curve

DWD: dimension-wise distance

EHR: electronic health record

GAN: generative adversarial network

ICD-9/10: International Classification of Disease, Ninth or Tenth Revision

MIMIC-IV: Medical Information Mart for Intensive Care, the Fourth Version

ML: machine learning

TRTR: training on real testing on real

TSTR: training on synthetic testing on real

Edited by K El Emam, B Malin; submitted 10.09.23; peer-reviewed by A Hashemi, C Sun; comments to author 16.10.23; revised version received 24.01.24; accepted 07.03.24; published 22.04.24.

Please cite as:

Yan C, Zhang Z, Nyemba S, Li Z

Generating Synthetic Electronic Health Record Data Using Generative Adversarial Networks: Tutorial

JMIR AI 2024;3:e52615

URL: <https://ai.jmir.org/2024/1/e52615>

doi: [10.2196/52615](https://doi.org/10.2196/52615)

PMID: [38875595](https://pubmed.ncbi.nlm.nih.gov/38875595/)

©Chao Yan, Ziqi Zhang, Steve Nyemba, Zhuohang Li. Originally published in JMIR AI (<https://ai.jmir.org/>), 22.04.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR AI, is properly cited. The complete bibliographic information, a link to the original publication on <https://www.ai.jmir.org/>, as well as this copyright and license information must be included.

Original Paper

Cost, Usability, Credibility, Fairness, Accountability, Transparency, and Explainability Framework for Safe and Effective Large Language Models in Medical Education: Narrative Review and Qualitative Study

Majdi Quttainah¹, PhD; Vinaytosh Mishra², PhD; Somayya Madakam³, PhD; Yotam Lurie⁴, PhD; Shlomo Mark⁵, PhD

¹College of Business Administration, Kuwait University, Kuwait, Kuwait

²College of Healthcare Management and Economics, Gulf Medical University, Ajman, United Arab Emirates

³Information Technology, Birla Institute of Management Technology, Knowledge Park - II, Greater Noida, India

⁴Department of Management, Ben-Gurion University, Negev, Israel

⁵Department of Software Engineering, Shamoon College of Engineering, Ashdod, Israel

Corresponding Author:

Vinaytosh Mishra, PhD

College of Healthcare Management and Economics

Gulf Medical University

Al Jurf 1

Ajman, 4184

United Arab Emirates

Phone: 971 503310560

Email: vinaytosh@gmail.com

Abstract

Background: The world has witnessed increased adoption of large language models (LLMs) in the last year. Although the products developed using LLMs have the potential to solve accessibility and efficiency problems in health care, there is a lack of available guidelines for developing LLMs for health care, especially for medical education.

Objective: The aim of this study was to identify and prioritize the enablers for developing successful LLMs for medical education. We further evaluated the relationships among these identified enablers.

Methods: A narrative review of the extant literature was first performed to identify the key enablers for LLM development. We additionally gathered the opinions of LLM users to determine the relative importance of these enablers using an analytical hierarchy process (AHP), which is a multicriteria decision-making method. Further, total interpretive structural modeling (TISM) was used to analyze the perspectives of product developers and ascertain the relationships and hierarchy among these enablers. Finally, the cross-impact matrix-based multiplication applied to a classification (MICMAC) approach was used to determine the relative driving and dependence powers of these enablers. A nonprobabilistic purposive sampling approach was used for recruitment of focus groups.

Results: The AHP demonstrated that the most important enabler for LLMs was *credibility*, with a priority weight of 0.37, followed by *accountability* (0.27642) and *fairness* (0.10572). In contrast, *usability*, with a priority weight of 0.04, showed negligible importance. The results of TISM concurred with the findings of the AHP. The only striking difference between expert perspectives and user preference evaluation was that the product developers indicated that *cost* has the least importance as a potential enabler. The MICMAC analysis suggested that cost has a strong influence on other enablers. The inputs of the focus group were found to be reliable, with a consistency ratio less than 0.1 (0.084).

Conclusions: This study is the first to identify, prioritize, and analyze the relationships of enablers of effective LLMs for medical education. Based on the results of this study, we developed a comprehensible prescriptive framework, named CUC-FATE (Cost, Usability, Credibility, Fairness, Accountability, Transparency, and Explainability), for evaluating the enablers of LLMs in medical education. The study findings are useful for health care professionals, health technology experts, medical technology regulators, and policy makers.

KEYWORDS

large language model; LLM; ChatGPT; CUC-FATE framework; cost, usability, credibility, fairness, accountability, transparency, and explainability; analytical hierarchy process; AHP; total interpretive structural modeling; TISM; medical education; adoption; guideline; development; health care; chat generative pretrained transformer; generative language model tool; user; innovation; data generation; narrative review; health care professional

Introduction

Background

Natural language programming solutions have been available for the last 15 years. However, these models recently witnessed an avalanche breakdown with the launch of ChatGPT by OpenAI, a company that was only established recently (December 2015) after receiving an investment from Elon Musk and others. ChatGPT is a generative language model tool that enables users to converse with machines about various subjects. With 1.6 billion monthly users, this freemium is the fastest-growing application in the history of the internet. Since its release on November 30, 2022, ChatGPT has sparked much discussion and enthusiasm in multiple industries, including medicine. ChatGPT and related technologies have been identified as disruptive innovations with the potential to revolutionize academia and scholarly publishing [1]. Additionally, preliminary research suggests that ChatGPT has practical applications throughout the clinical workflow [2].

The introduction of ChatGPT and the subsequent release of several extended products and functional plugins have profoundly impacted scientific researchers. These products have also influenced the ideas and methodologies used in traditional research, including recommendation, emotion recognition, and information generation. ChatGPT's assistance has improved some of the associated work in these fields, particularly with providing helpful supplementary information to raise the caliber of data generation. With the integration of machine learning and artificial intelligence (AI) technologies, medical imaging has advanced quickly. Among these developments, using cutting-edge language models such as large language models (LLMs), ChatGPT, and GPT-4 has shown significant promise in elevating several elements of medical imaging and revolutionizing radiology. These models can produce and comprehend human-like text owing to access to various textbooks, journals, and research materials available on the internet. This could provide the necessary context and prior knowledge to support a variety of tasks involving medical imaging, such as synthesis, reconstruction, analysis, segmentation, interpretation, automated reporting, and more. These technologies have further been improved using supervised and reinforcement learning methods based on OpenAI's GPT LLMs. These models have shown excellent performance in various natural language processing (NLP) tasks, including language translation, text summarization, and question-answering. The models have been pretrained on enormous amounts of text data. Users can ask questions, obtain responses, and engage in genuine conversation with the bot given ChatGPT's human-like conversational experience.

ChatGPT and other LLMs remain a research hotspot in multimedia analysis and application. However, several crucial difficulties must be resolved, including (1) improving interactions with ChatGPT to collect more useful auxiliary information, (2) methods to combine ChatGPT with traditional inquiries to fully exploit its benefits, and (3) analyzing the data obtained from ChatGPT for their incorporation with the intended usage. A particularly significant challenge is to effectively use past information obtained with such huge models and to ensure consistency and complementary features across many modalities to improve multimodal generation performance, which is especially relevant for AI-generated content. The finest use cases for ChatGPT, a well-liked chatbot built on a potent AI language model, are still being worked out. ChatGPT can provide help in writing an essay, thesis, or dissertation by creating a research question, developing a plan, developing literary concepts, rewriting text, and getting feedback. Moreover, the NLP and automated data analysis capabilities offered by ChatGPT enable researchers, marketers, and organizations to analyze text quickly and accurately. Via its AI-powered functions, ChatGPT can help to spot significant trends and insights in a data set that might otherwise be challenging to find. Additionally, ChatGPT can assist with the creation of top-notch prompts for paper analysis.

LLM Functionality

ChatGPT is a prediction system that anticipates what it should write based on previously processed texts. This type of AI is known as a language model. However, ChatGPT offers more promise than its predecessors given that it is trained on enormous amounts of data, with the majority of these data originating from the abundant supply of data available on the internet. According to OpenAI, ChatGPT was also trained on examples of back-and-forth human interaction, which results in a conversation style that is much more human than that of other chatbots, thus advancing the capability of NLP solutions.

NLP is a field of AI employing linguistics, statistics, and machine learning to enable computers to comprehend spoken language. NLP systems can infer meaning from spoken or written words, including all of the subtleties and complexities of an accurate narrative text. This makes it possible for machines to obtain value from even unstructured data. NLP has witnessed significant advancements in recent years. An LLM is a deep-learning algorithm that can be used to perform NLP tasks, including, among other abilities, summarizing and generating text. As one of the main applications, LLM-based chatbots are computer programs that can simulate conversations with human users. NLP techniques can be used to enable chatbots to understand and respond to user input. LLM uses deep-learning techniques to understand and generate human language, which requires training on vast amounts of text data and then uses

statistical algorithms to learn patterns and relationships within language. These models can perform various tasks, including language translation, question-answering, sentiment analysis, and summarization. With ChatGPT, users can learn, compare, and validate answers for different academic subjects, including physics, math, and chemistry, as well as abstract topics such as philosophy and religion [3]. Users can also generate human-like text such as news articles, chatbot conversations, and even literary works such as essays and romantic poems. The main difference of GPTs from other LLMs lies in their architecture and training methodology. GPTs are based on a deep-learning architecture known as a “transformer.” Transformers are designed to process sequential data such as language more efficiently than other architectures. LLMs are currently at the forefront of intertwining AI systems with human communication and everyday life [4]. Large pretrained language models have significantly advanced NLP research with respect to various applications [5,6]. Although these more complicated language models can produce complex and coherent natural language, several recent studies have shown that they can also pick up unfavorable social biases that can feed into negative stereotypes [7].

NLP in Health Care

Health care consumers may turn to the research literature for information not provided in patient-friendly documents. However, reading medical literature can be difficult. One study identified four key elements made possible by NLP to increase access to medical papers: explanations of foreign terminology, plain language section summaries, a list of crucial questions that direct readers to the portions that provide the answers, and simple language summaries of those passages [8]. Significant advancements in smart health care have been made in recent years, with new AI technologies enabling a range of intelligent applications in various health care contexts. NLP, as a fundamental AI-powered technology that can analyze and comprehend human language, is crucial for smart health care [9]. NLP methods have been utilized to organize data in health care systems by sifting out pertinent information from narrative texts to offer information for decision-making. Thus, NLP approaches help to lower health care costs and are essential for streamlining health care procedures [10]. Advancements in NLP will make robotic process automation possible in health care, which can further drive efficiency. Health care data are complex, which should be given due consideration at the time of designing health care applications. Deep-learning approaches such as convolutional neural network and recurrent neural network models have become prominent in health care applications, demonstrating promising accuracy. Nevertheless, there is still substantial room for improvement of these models to enable their usage without human supervision. Deep-learning techniques offer an effective and efficient model for data analysis by revealing hidden patterns and extracting valuable information from a large volume of health data, which standard analytics cannot perform within a given time frame [11].

ChatGPT in Medical Education

ChatGPT has many potential applications in health care education, research, and practice [12], which can enhance

medical education by helping students develop subjective learning and expression skills [13]. The number of ChatGPT users has shown exponential growth and the tool is increasingly utilized by students, residents, and attending physicians to direct learning and answer clinical questions [14]. However, authors using ChatGPT professionally for academic work should exercise caution as it remains unclear how ChatGPT handles hazardous content, false information, or plagiarism [15]. While ChatGPT can simplify the task of radiological reporting, there is still a chance of inaccurate statements and missing medical information [15]. Therefore, the tool needs refinement before it can be used widely with confidence in medicine [16]. A recent review explored ChatGPT’s applications and reported various challenges such as ethical concerns, data biases, and safety issues [17]. Thus, it is imperative to balance AI-assisted innovation and human expertise [18]. ChatGPT has quickly gained significant attention from academia, research, and industries despite these shortcomings. The first aim of this study was therefore to determine the requirements, or enablers, for a successful LLM application in medical education using a narrative review of the existing literature.

Enablers of LLM for Medical Education

For the purpose of this study, we refer to enablers as the factors, resources, or conditions that facilitate or support achieving a good LLM application for medical education. Medical education prepares would-be physicians and other health care professionals with the knowledge, skills, and attitudes necessary for competent and compassionate patient care. The general definition of an enabler is a factor that makes it easier for a goal to be realized or for someone to accomplish a particular task. Enablers of LLM for medical education can be tangible or intangible and should play a crucial role in achieving the outcomes expected from the application.

As LLMs are trained on massive data, they are resource-demanding tools. Therefore, the cost of training an LLM for medical education may be prohibitive [19]. Accordingly, it is imperative to use efficient computing to address this issue [20]. Usability is one of the key criteria that determines the usefulness of an application in medical education, and LLMs are no exception [21]. The extant literature has highlighted usability as an important criterion for the successful implementation of a new technology in education [22]. Similarly, the credibility of an application is another very important factor for technological interventions used in medical education [23,24]. Although ChatGPT has disclaimers about the source of information provided, it does not disclose its sources categorically, and can sometimes hallucinate about the source, which may be misleading to the user. LLMs also have reported issues with fairness, computation, and privacy. By perpetuating social prejudices and stereotypes, they risk causing unfair discrimination and physical harm, along with potential harm to the user’s reputation [25]. Ma et al [26] provided an overview of fairness of LLMs in multilingual and non-English situations, emphasizing the limitations of recent studies and the challenges faced by English-only methodologies [26].

Another issue of LLMs such as ChatGPT is related to their accountability, generally defined as taking responsibility for

one's obligation to treat others honestly and morally. However, it is unclear who will be held accountable and responsible if the LLM provides incorrect recommendations or forecasts for a particular downstream activity. Overall, employing LLMs is associated with considerable risk; therefore, precautions must be taken to minimize these risks and ensure their ethical and responsible use. To foster a cross-disciplinary global inclusive consensus on the ethical use, disclosure, and proper reporting of generative AI models such as GPT and other LLM technologies in academia, Cacciamani et al [23] proposed the ChatGPT, Generative Artificial Intelligence, and Natural Large Language Models for Accountable Reporting and Use Guidelines initiative in 2023. However, the underlying model of GPT3.5 deviates from the ethical guidelines proposed by Cacciamani et al [23]. Another important criterion reported for the medical applications of LLMs is transparency, which is an essential ethical consideration in the fields of science, engineering, business, and the humanities. Transparency refers to functioning in a way that makes it simple for others to observe what actions have been taken [27], thus representing a sign of responsibility, honesty, and openness. Conversely, LLMs are opaque to users. Recently suggested explainability techniques aim to make LLMs more transparent. Although these techniques are not a cure-all, they might form the basis for the development of models with fewer flaws or, at the very least, the ability to explain their logic. In their systematic experiments with synthetic data, Wu et al [28] demonstrated that autoregressive and masked language models can successfully learn to emulate semantic relations between expressions with strong transparency, where all expressions have context-independent denotations.

Finally, the LLMs used in medical education must be explainable, and the best freely available options lag in this respect. Most LLMs are complex models built using deep learning [29]; therefore, these models can produce better predictions with more information or network parameters, which comes at a cost of sacrificing explainability. Some models fail to describe how they came to their conclusion. Recently suggested explainability techniques aim to make language models more transparent. Even though these are not complete solutions, they can act as the basis for the development of less problematic models or, at the very least, models that can explain their logic. However, Du et al [30] identified false patterns detected by LLMs using explainability in their study.

Need for This Study

The need for this study arises from the rapid integration of LLMs such as ChatGPT in various fields, including medical education. Although LLMs offer promising benefits for health care, their effective integration in medical education remains a developing area. Accordingly, the aim of this study was to identify and prioritize the key enablers for successful LLM implementation in medical education. This can in turn help to address the lack

of comprehensive frameworks guiding the development and use of LLMs in this field. By exploring the dynamics of various enablers such as credibility, accountability, fairness, cost, usability, transparency, and explainability, this study provides a structured approach to enhance the quality and effectiveness of LLMs in educating health care professionals.

Specifically, this study was based on the following three major research questions: (1) What are the enablers of a suitable LLM application for medical education? (2) What is the relative importance of these enablers in achieving the goals of medical education? and (3) What is an approach to developing an LLM to achieve medical education goals? With this background, the following research objectives were set: (1) identify the enablers of a suitable LLM for medical education, (2) prioritize the identified enablers in achieving the goals of medical education, and (3) propose a framework for developing an LLM to achieve the medical education goals.

Methods

Study Design

To achieve the first research objective, we performed a narrative review of the extant literature published on technology solutions in medical education. A narrative review is a scholarly article synthesizing existing research on a particular topic in a narrative or story-like manner. Unlike systematic reviews or meta-analyses, which use rigorous methodologies to analyze and summarize research findings quantitatively, narrative reviews provide a qualitative, comprehensive overview of a subject. Narrative reviews often involve critical analysis and discussion, integrating the authors' expertise and interpretation. Narrative reviews are thus useful for obtaining a broad understanding of a topic and identifying trends, gaps, and controversies within a field.

Two authors (SM and VM) searched the Scopus, Web of Science, and Google Scholar databases to identify suitable literature for our narrative review. The inclusion criteria were articles published in the English language in the last 5 years. In the second stage, duplicates and articles for which the full text was unavailable were eliminated. The identified enablers from this review were then used to address the first research question. These enablers were presented in front of a focus group comprising seven experts working in universities and institutions delivering medical education in India and the United Arab Emirates to validate the selection (Table 1). The focus group endorsed the choice of the enablers for further research; in addition, one article published in 2010 was added on the recommendation of the focus group as it was found to be useful in explaining competing interests in medical education. One author (VM) facilitated the focus group discussion to obtain the finalized list of enablers.

Table 1. Characteristics of the focus group for validation of identified enablers.

Expert	Qualification	Experience (years)	Age (years)	Nationality
Cardiologist	Masters in Medicine	12	42	India
Endocrinologist	Masters in Medicine	20	45	India
Technology expert	Doctor of Philosophy	15	50	United Arab Emirates
Dentistry educator	Masters in Dentistry	10	40	United Arab Emirates
Podiatrist educator	Doctor of Philosophy	10	35	United Arab Emirates
Diabetes educator	Doctor of Philosophy	18	43	India
Nursing educator	Doctor of Philosophy	15	41	United Arab Emirates
Radiologist	Doctor of Philosophy	12	41	India

Analytical Hierarchy Process Modeling

An analytical hierarchy process (AHP) was utilized to achieve the second study objective of prioritizing the identified enablers for developing an LLM for medical education. The AHP is a popular method for determining the relative importance of the criteria in a multicriteria decision analysis task. To date, the AHP has been extensively used in the management and social science fields [31]. The advantage of this process is that it incorporates the mechanisms to assure reliability in the decision-making case of ambiguity. Some researchers have suggested using a “fuzzy” version of the AHP [32] and others have suggested using the entropy weight method to reduce the negative effect of individual subjective evaluation bias on the accuracy of comprehensive evaluation [33]. Since the ranking obtained by the AHP method was further validated by total interpretive structural modeling (TISM) in this study (see below), fuzzy logic or entropy weight was avoided in our AHP modeling. The five steps used for AHP are: (1) defining the decision problem, (2) creating a hierarchy, (3) pairwise comparison, (4) deriving a weighted priority, and (5) consistency check for decision. We used the Delphi method for pairwise comparisons. A cut-off value of 75% was used to accept the value for the pairwise comparison. The standard scale proposed by Saaty [34] was used for the pairwise comparison.

TISM and Focus Groups

Finally, to address the third research objective, we investigated the relationships among key enablers to inform the development

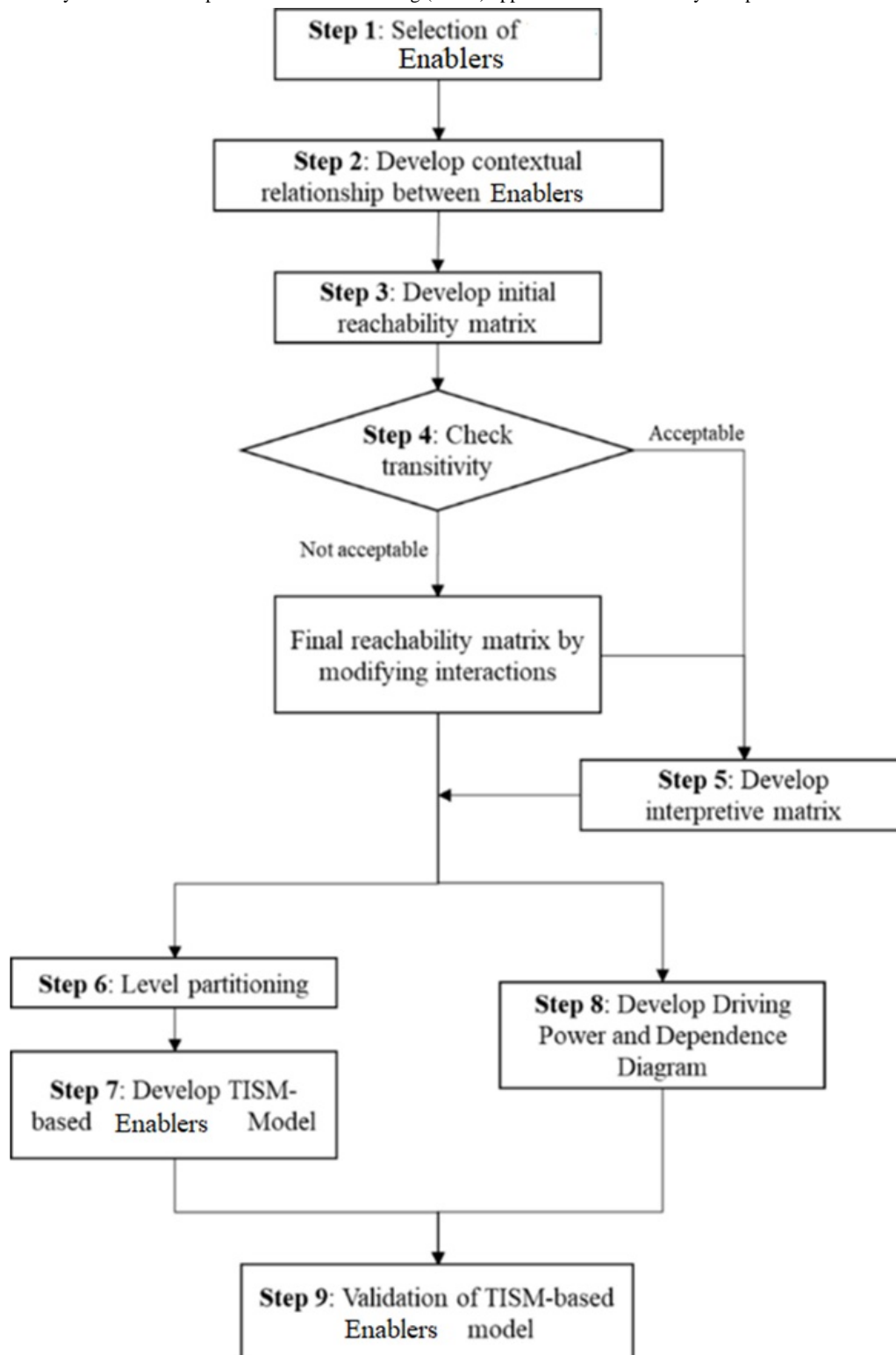
of a suitable medical education LLM. A qualitative research design is useful to understand a phenomenon under study rather than assessing the strength and direction of causal relationships in a conceptual model [35]. For this purpose, we established a focus group with five experts in the fields of information technology and product development with relevant research experience. The details of this expert group are provided in [Table 2](#).

According to the information obtained from the focus group, TISM was used to model the enablers for a medical education LLM application. In his seminal paper, Sushil [36] provides a detailed account of the interpretation of interpretive structural modeling and TISM, highlighting the advantage of the latter over the former. For the sake of brevity, we have not included the details of the TISM method herein, which can be found in the relevant literature [37]. In brief, TISM is a process that converts poorly articulated mental models of systems into visible and well-defined models that are useful for gaining better understanding and decision-making. The presence and absence of a relationship between enablers were ascertained based on an unstructured interview of the focus group conducted by one researcher (SM). If more than 50% of the focus group members indicated that there is a relationship between two enablers, the enabler was considered to be present, which was coded as “Y.” An overview of the TISM approach used in this study is provided in [Figure 1](#).

Table 2. Characteristics of the focus group used for total interpretive structural modeling.

Expert	Qualification	Experience (years)	Age (years)	Country
Product development	Masters in management	21	42	Singapore
Product development	Bachelors in engineering	21	42	United Arab Emirates
Technology expert	Bachelors in engineering	19	40	India
Technology expert	Masters in engineering	10	33	India
Decision science expert	Doctor of Philosophy	10	38	India

Figure 1. Summary of the total interpretive structural modeling (TISM) approach used in the study. Adapted from Mishra and Rana [33].



We further used cross-impact matrix multiplication applied to classification (MICMAC) analysis to evaluate the direct and indirect relationships among various elements in a complex system. MICMAC analysis is applied to the reachability matrix to classify the elements into four categories based on their driving power (ability to influence other elements) and dependence (level of being influenced by other elements).

Ethical Considerations

This study, involving a qualitative focus group discussion, did not require approval from an ethical review board as it did not involve human subjects in a manner necessitating such review. No informed consent was required for the same reason. However, to maintain ethical standards, we ensured that all data collected were either anonymized or deidentified. This means that any information that could potentially identify individual participants was removed or altered to protect their privacy. No

compensation was provided to participants, as is common in studies of this nature. This decision was made considering the study design and the ethical imperative to avoid undue influence on participants' responses. The absence of compensation was communicated to all participants. Throughout the study, we adhered to strict data protection protocols to safeguard the confidentiality of the information shared during the focus group discussions. These measures included secure data storage, restricted access to authorized personnel, and adherence to data protection laws and regulations. This approach ensured that the privacy and integrity of participant information were always maintained.

Results

AHP Modeling

Based on the selected enablers identified for developing a suitable LLM medical education application according to the narrative review of the literature (Table 3), the focus group was asked to provide their input for pairwise comparison, and the resultant matrix [A] is presented in Table 4.

Once the initial comparison matrix was determined, the matrix was normalized and an average of each row was taken to calculate the priority weight [X]. The normalized matrix, priority weight, and rank of the enablers are given in Table 5. The priority weight, as the eigenvector, was further used to calculate the consistency ratio (CR).

Table 3. Summary of reported enablers of large language models for medical education.

Enabler code	Enabler	Description	References
E1	Cost	Cost of computation, including hardware, software, and energy requirement	[19,20]
E2	Usability	User-centric design, ease of use, and positive user experiences	[21,22]
E3	Credibility	Level of trust and reliability that users place in the application	[23,24]
E4	Fairness	Absence of unfair discrimination, physical harm, and harm to user reputation	[25,26]
E5	Accountability	Taking responsibility for the obligation to treat users with honesty and morality	[27,38]
E6	Transparency	Functioning in a way that makes it simple for others to observe what actions are taken	[27,30]
E7	Explainability	Ability to describe how the models came to their conclusion	[29,30]

Table 4. Initial pairwise comparison matrix for the analytical hierarchy process.a

Enablers	Cost (E1)	Usability (E2)	Credibility (E3)	Fairness (E4)	Accountability (E5)	Transparency (E6)	Explainability (E7)
E1	1	3	0.2	1	0.2	3	3
E2	0.33	1	0.11	0.33	0.11	1	1
E3	5	9	1	5	5	3	3
E4	1	3	0.2	1	0.2	3	3
E5	5	9	0.2	5	1	5	5
E6	0.33	1	0.33	0.33	0.2	1	1
E7	0.33	1	0.33	0.33	0.2	0.2	1

^aNumbers represent the pairwise comparison of different enablers using the scale developed by Saaty [34].

Table 5. Normalized matrix and priority weight of enablers.

Enablers	Cost (E1)	Usability (E2)	Credibility (E3)	Fairness (E4)	Accountability (E5)	Transparency (E6)	Explainability (E7)	Priority weight	Rank
E1	0.077	0.1111	0.0844	0.077	0.0289	0.1852	0.1765	0.10572	3
E2	0.0254	0.037	0.0464	0.026	0.0159	0.0617	0.0588	0.03871	7
E3	0.3849	0.3333	0.4219	0.385	0.7236	0.1852	0.1765	0.37289	1
E4	0.077	0.1111	0.0844	0.077	0.0289	0.1852	0.1765	0.10572	3
E5	0.3849	0.3333	0.0844	0.385	0.1447	0.3086	0.2941	0.27642	2
E6	0.0254	0.037	0.1392	0.025	0.0289	0.0617	0.0588	0.0538	5
E7	0.0254	0.037	0.1392	0.025	0.0289	0.0123	0.0588	0.04674	6

Based on this matrix, the eigenvector X was calculated according to the following equation:

$$[A] X = \lambda_{max} X - (1)$$

Using the data in Tables 4 and 5, λ_{max} was obtained as follows:

$$[A]X = [0.76, 0.28, 3.46, 0.76, 2.26, 0.39, 0.34] - (2)$$

$$\lambda_{max} = \text{average} \{0.76/0.11, 0.24/0.04, 3.46/0.37, 0.76/0.11, 0.39/0.05, 0.34/0.05\} - (3)$$

$$\lambda_{max} = 7.66 - (4)$$

The consistency index (CI) was then calculated based on the λ_{max} as follows: $CI = (7.66 - 7)/6 = 0.11 - (5)$. Finally, the CR of the judgment was calculated by dividing the CI by the random index (RI). The RI value for a 7×7 matrix is 1.32 from the RI table. Thus, the CR becomes 0.084; as this is less than 0.1, it is considered to be acceptable.

Modeling Relationships Among Enablers

We further used TISM for ascertaining the relationships among these seven enablers. Table 6 shows a matrix indicating the interrelationships between the enablers listed in Table 3, with “Y” indicating the existence of a relationship and “N” indicating no relationship. The resultant matrix is referred to as the structural self-interaction matrix.

In the next step, we replaced all “Ys” with 1s and all “Ns” with 0s and incorporated the transitivity rule to obtain the final reachability matrix shown in Table 7.

The next step in developing LLMs for medical education involved listing reachability and antecedent sets for each enabler, followed by level partitioning, which is an iterative process of assigning enablers at different levels. Enablers with similar intersection sets as reachability sets are placed at the top level. The process is then repeated until levels are established for all enablers. In this study, all enablers were assigned after three iterations; hence, there are three levels in the hierarchy. The summary of level partitioning is provided in Table 8. The level of an enabler is a reflection of its driving power and dependence power, as indicated in Table 7. The higher the level of the enabler, the more dependent it is, whereas the driving ability improves when moving to lower levels.

Once the level partitioning was complete, the TISM was developed and presented to the focus group for validation. Only significant transitive links were included in the model to facilitate interpretation. The final digraph for the TISM developed in the study is depicted in Figure 2.

Table 6. Structural self-interaction matrix for the identified enablers of large language models for medical education.

Enablers	Cost (E1)	Usability (E2)	Credibility (E3)	Fairness (E4)	Accountability (E5)	Transparency (E6)	Explainability (E7)
E1	Y ^a	Y	N ^b	N	N	Y	N
E2	Y	Y	N	N	N	Y	Y
E3	N	N	Y	Y	Y	N	N
E4	N	N	Y	Y	N	N	N
E5	N	N	Y	N	Y	N	N
E6	Y	Y	N	N	N	Y	Y
E7	N	Y	N	N	N	Y	Y

^aY: existence of a relationship between two enablers.

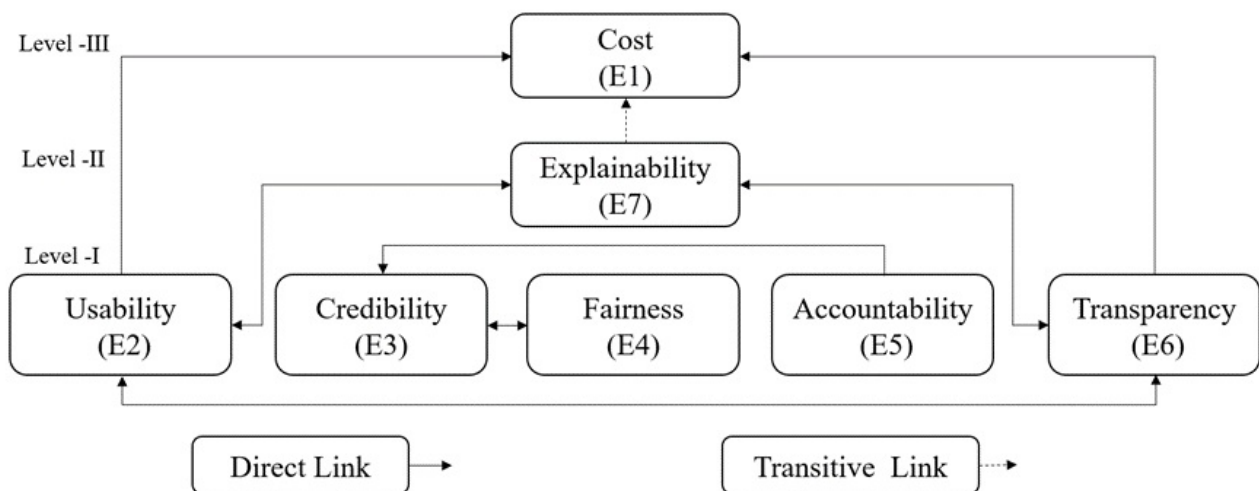
^bN: no relationship exists between two enablers.

Table 7. Final reachability matrix of the enablers for developing large language models in medical education.

Enablers	Cost (E1)	Usability (E2)	Credibility (E3)	Fairness (E4)	Accountability (E5)	Transparency (E6)	Explainability (E7)	Driving power
E1	1	1	0	0	0	1	1	4
E2	1	1	0	0	0	1	1	4
E3	0	0	1	1	1	0	0	3
E4	0	0	1	1	0	0	0	2
E5	0	0	1	0	1	0	0	2
E6	1	1	0	0	0	1	1	4
E7	0	1	0	0	0	1	1	3
Dependence power	3	4	3	2	2	4	4	Not applicable

Table 8. Summary of label partitioning iterations (1 to 6).

Enablers, (Mi)	Reachability set, R(Mi)	Antecedent set, A(Ni)	Intersection set, R(Mi)∩A(Ni)	Level
1	1	1	1	III
2	1, 2, 6, 7	1, 2, 6, 7	1, 2, 6, 7	I
3	3, 4, 5	3, 4, 5	3, 4, 5	I
4	3, 4	3, 4	3, 4	I
5	3, 5	3, 5	3, 5	I
6	1, 2, 6, 7	1, 2, 6, 7	1, 2, 6, 7	I
7	7	1, 7	7	II

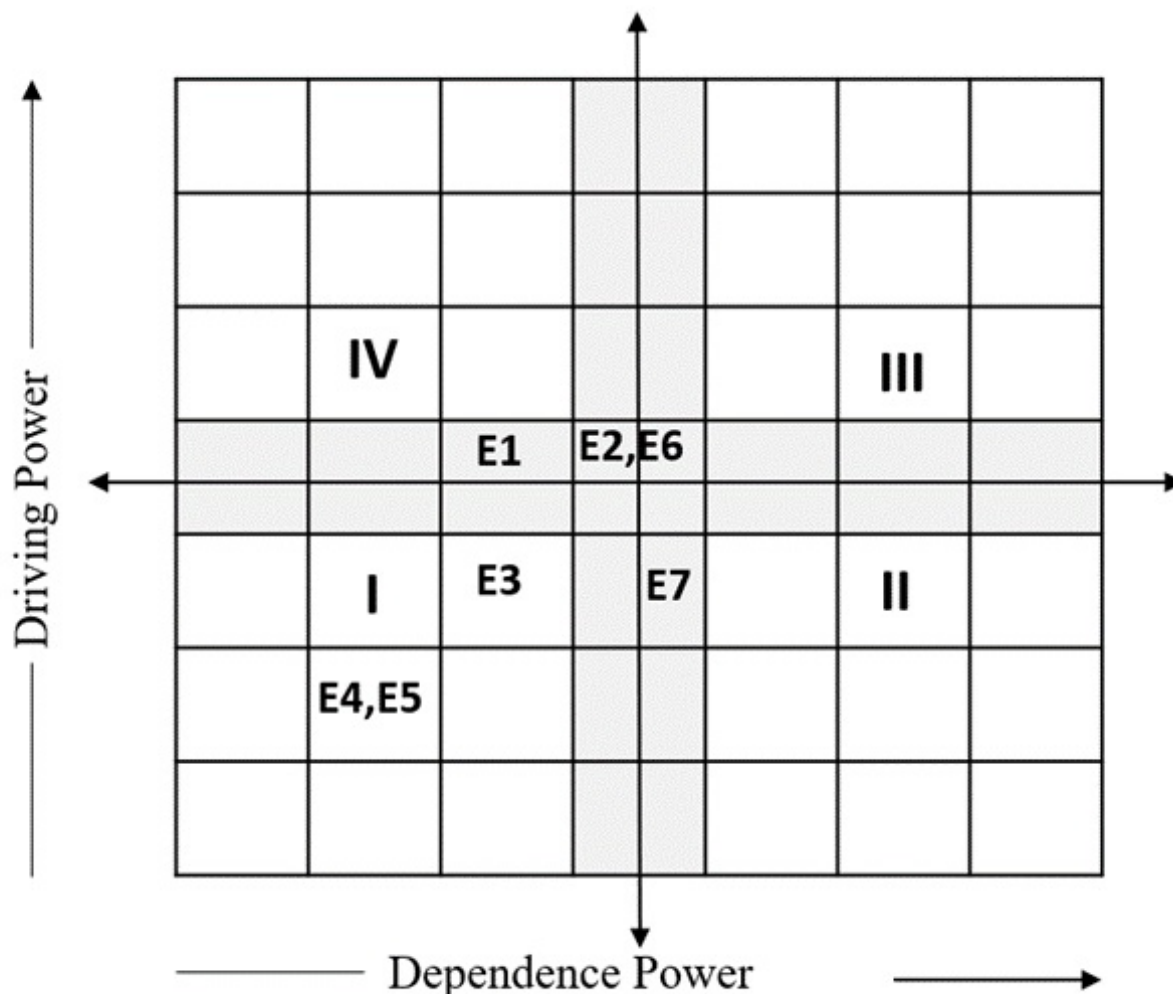
Figure 2. Diagraph of the total interpretive structural model for the development of large language models in medical education.

Validation Analysis

We further used MICMAC analysis to validate the study findings and derive conclusions. MICMAC analysis involves the development of a graph that classifies enablers based on their driving and dependence power. As shown in Figure 3, the first quadrant contains autonomous enablers E3 (Credibility), E4 (Fairness), and E6 (Accountability), indicating that the

variables falling in this quadrant have low driving and dependence powers. The two enablers falling in the grey region between the third (linkage) and fourth (independent) quadrants are E2 (Usability) and E6 (Transparency), which have medium driving and dependence powers. Similarly, E7 (Explainability) falls in the grey region between the first (autonomous) and second (dependent) variables. Finally, E1 (Cost) falls under the fourth (independent) quadrant.

Figure 3. Cross-impact matrix-based multiplication applied to a classification (MICMAC) analysis for enablers of a large language model in medical education. I-IV indicate different levels of the enablers E1-E7. E1: cost; E2: usability; E3: credibility; E4: fairness; E5: accountability; E6: transparency; E7: explainability.



Discussion

Principal Findings

The results of the AHP suggested that credibility, followed by accountability are the foremost enablers for effective LLMs in medical education. The extant literature supports this finding, in highlighting the relevance of the source of information based on which the response was generated [39]. Similarly, the importance of defining accountability has been emphasized in the recent literature. For example, Tan et al [40] advocate for accountability as an important factor in increasing the adoption of LLMs in medical education, training, and practice. The next most important factors to consider are ethical issues such as fairness and cost. LLMs have been criticized for bias against gender or ethnic groups [17]. These problems need to be addressed to make LLMs effective in medical education. Moreover, training LLMs on billions of parameters is demanding; thus, only technology giants will launch these LLMs [41]. Governments should therefore ensure that the cost of using these LLMs does not become prohibitive for end users, who may resort to insufficient solutions that could ultimately affect the safety of patients.

In contrast to existing studies, transparency and explainability ranked fifth and sixth in importance in our analysis [40]. Many best practices related to health technology suggest that models should use explainable AI in medical devices [17]. The low priority of these enablers identified in this study indicates that the end user is unaware of the criticality of these factors; thus, health care professionals need to be educated about these issues as they are not technology savvy [42]. Governments should also establish guidelines for the approval of Software as Medical Devices so that these enablers are taken care of at the product development stage. Finally, the focus group indicated that usability is the least important factor among the seven enablers discussed. Although general-purpose LLMs such as ChatGPT are less cluttered, their performance is input-dependent. Improving the prompt use of the recommendation system can enhance the usability and accuracy of LLMs in medical education [43]. The expert group advised that the LLMs will improve on these factors with time.

The results from TISM suggested a slight difference in the perspective of product developers and end users, as the experts gave equal importance to the enablers credibility, fairness, accountability, transparency, and explainability. These results are consistent with extant literature published in peer-reviewed

journals [40,41], as these are all features related to model development and training.

In contrast to earlier studies, the product developers and technology experts placed less significance on usability as an enabler, which was given a medium level [43]. Thus, the finding of the TISM validates the results of the AHP. The only difference was that cost was considered as the least important enabler for product developers. However, a recent study indicated that economic and environmental costs are significant factors in developing general-purpose LLMs [44].

Successful LLM development involves a complex interplay among technical innovation, regulatory compliance, production costs, and end-user needs. The aim should be to develop products that excel in functionality and positively impact the lives of those who rely on them without causing financial hardship. Thus, this study calls for collaboration between product developers, original equipment manufacturers, regulators, and other stakeholders to find solutions that align with technological advancements and societal expectations for affordability and accessibility.

Finally, the findings of this study were validated using MICMAC analysis, creating a graph that categorizes enablers based on their driving power and dependence power. In this graph, the enablers credibility, fairness, and accountability are in the first quadrant (autonomous) with low power, indicating that these variables are relatively independent and have limited influence on other variables. Usability and transparency are in the grey region between the third (linkage) and fourth (independent) quadrants with medium power, indicating a moderate influence on other variables and similarly influenced by them. Explainability falls in the grey region between the first (autonomous) and second (dependent) quadrants, also indicating a medium influence on other variables and a similar influence on them. Finally, cost falls under the fourth quadrant (independent), suggesting that it strongly influences other enablers without being significantly influenced by them. MICMAC analysis comprehensively explains the relationships and dynamics among variables within a complex system. This can help decision makers identify key drivers, dependencies, and interactions, enabling them to make informed strategic decisions and allocate resources effectively.

Acknowledgments

The authors are highly indebted to all focus group participants for their time and effort. The authors are also obliged to their respective institutions for the infrastructural support provided. The authors disclose using the artificial intelligence tools Grammarly and Quillbot for manuscript language editing. The article processing charges for the publication of the manuscript are funded by the College of Business Administration, Kuwait University.

Data Availability

The necessary data and calculations for the analytic hierarchy process model and the self-interaction matrix for the total interpretive structural model are available on a GitHub repository [46].

Practical and Theoretical Implications

The study has one implication each for theory and for practice. For theory, this study extends the Fairness, Accountability, Transparency, and Explainability (FATE) framework [45] into a more comprehensive Cost, Usability, Credibility, Fairness, Accountability, Transparency, and Explainability (CUC-FATE) framework for developing LLMs for health care professionals. With respect to the implication for practice, this study is the first of its kind and provides a prescriptive framework for developing LLMs in health care, especially medical education. The findings of this study are useful for policy makers, medical device regulators, education policy makers, health care professionals, and product developers at the helm of creating Software as a Medical Device.

Limitations

One of the limitations of the study is that the results largely rely on experts from India and the United Arab Emirates. Although technology and health care practices are standardized globally, the findings should only be generalized to the populations from these regions. This study provides insight into the relationships between different enablers but we did not further evaluate the strength of these associations. Graph theory or structured equation modeling can be used to address these gaps in future studies.

Conclusion

This study emphasizes key factors for effective LLMs in medical education: credibility and accountability are vital enablers, while addressing bias and cost is crucial for enhancing LLM potential. Although important, transparency and explainability rank lower as LLM enablers among health professionals, suggesting a need for further education on this technology. Usability emerged as the least important factor; however, enhancing prompt use improves LLM accuracy. This study highlights a slight difference between product developers and end users. Although both groups prioritize credibility, fairness, accountability, transparency, and explainability, usability ranks lower for developers. Successful LLM development must balance innovation, compliance, costs, and user needs. Collaboration among stakeholders is crucial for aligning with technology and societal expectations.

Authors' Contributions

Conceptualization: VM, MQ, S Madkam, YL, and S Mark; Data curation: VM, S Madkam; Formal Analysis: VM, YL, and S Mark; Funding acquisition: MQ; Methodology: VM, MQ; Project administration: MQ; Supervision: YL and S Mark; Validation: YL and S Mark; Visualization: VM; Writing—original draft: VM, MQ; Writing—review & editing: YM and S Mark.

Conflicts of Interest

None declared.

References

1. Haque MUI, Dharmadasa I, Sworna ZT, Rajapakse RN, Ahmad H. "I think this is the most disruptive technology": exploring sentiments of ChatGPT early adopters using Twitter data. arXiv. 2022. URL: <https://arxiv.org/abs/2212.05856> [accessed 2023-12-20]
2. Nastasi A, Courtright KR, Halpern SD, Weissman GE. A vignette-based evaluation of ChatGPT's ability to provide appropriate and equitable medical advice across care contexts. *Sci Rep* 2023 Oct 19;13(1):17885. [doi: [10.1038/s41598-023-45223-y](https://doi.org/10.1038/s41598-023-45223-y)] [Medline: [37857839](https://pubmed.ncbi.nlm.nih.gov/37857839/)]
3. Liu Y, Han T, Ma S, Zhang J, Yang Y, Tian J, et al. Summary of ChatGPT-related research and perspective towards the future of large language models. *Meta-Radiology* 2023 Sep;1(2):100017. [doi: [10.1016/j.metrad.2023.100017](https://doi.org/10.1016/j.metrad.2023.100017)]
4. Hagendorff T. Machine psychology: investigating emergent capabilities and behavior in large language models using psychological methods. arXiv. 2023 Mar. URL: <https://arxiv.org/abs/2303.13988> [accessed 2023-12-20]
5. Májovský M, Černý M, Kasal M, Komarc M, Netuka D. Artificial intelligence can generate fraudulent but authentic-looking scientific medical articles: Pandora's box has been opened. *J Med Internet Res* 2023 May 31;25:e46924 [FREE Full text] [doi: [10.2196/46924](https://doi.org/10.2196/46924)] [Medline: [37256685](https://pubmed.ncbi.nlm.nih.gov/37256685/)]
6. Brown T, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, et al. Language models are few-shot learners. 2020 Presented at: NIPS'20: 34th International Conference on Neural Information Processing Systems; December 6-12, 2020; Vancouver, BC.
7. May C, Wang A, Bordia S, Bowman SR, Rudinger R. On measuring social biases in sentence encoders. arXiv. 2019. URL: <https://arxiv.org/abs/1903.10561> [accessed 2023-12-20]
8. August T, Wang LL, Bragg J, Hearst MA, Head A, Lo K. Paper plain: Making medical research papers approachable to healthcare consumers with natural language processing. *ACM Trans Comput Hum Interact* 2023 Sep 23;30(5):1-38. [doi: [10.1145/3589955](https://doi.org/10.1145/3589955)]
9. Kaelin VC, Valizadeh M, Salgado Z, Parde N, Khetani MA. Artificial intelligence in rehabilitation targeting the participation of children and youth with disabilities: scoping review. *J Med Internet Res* 2021 Nov 04;23(11):e25745 [FREE Full text] [doi: [10.2196/25745](https://doi.org/10.2196/25745)] [Medline: [34734833](https://pubmed.ncbi.nlm.nih.gov/34734833/)]
10. Iroju OG, Olaleke JO. A systematic review of natural language processing in healthcare. *Int J Inf Technol Comput Sci* 2015 Jul 08;7(8):44-50 [FREE Full text] [doi: [10.5815/ijitcs.2015.08.07](https://doi.org/10.5815/ijitcs.2015.08.07)]
11. Lavanya P, Sasikala E. Deep learning techniques on text classification using natural language processing (NLP) in social healthcare network: a comprehensive survey. 2021 Presented at: 3rd International Conference on Signal Processing and Communication (ICPSC); May 13-14, 2021; Coimbatore, India. [doi: [10.1109/icspc51351.2021.9451752](https://doi.org/10.1109/icspc51351.2021.9451752)]
12. Sallam M. ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. *Healthcare* 2023 Mar 19;11(6):887 [FREE Full text] [doi: [10.3390/healthcare11060887](https://doi.org/10.3390/healthcare11060887)] [Medline: [36981544](https://pubmed.ncbi.nlm.nih.gov/36981544/)]
13. Seetharaman R. Revolutionizing medical education: can ChatGPT boost subjective learning and expression? *J Med Syst* 2023 May 09;47(1):61. [doi: [10.1007/s10916-023-01957-w](https://doi.org/10.1007/s10916-023-01957-w)] [Medline: [37160568](https://pubmed.ncbi.nlm.nih.gov/37160568/)]
14. Grabb D. ChatGPT in medical education: a paradigm shift or a dangerous tool? *Acad Psychiatry* 2023 Aug;47(4):439-440. [doi: [10.1007/s40596-023-01791-9](https://doi.org/10.1007/s40596-023-01791-9)] [Medline: [37160840](https://pubmed.ncbi.nlm.nih.gov/37160840/)]
15. Kleebayoon A, Wiwanitkit V. ChatGPT in medical practice, education and research: malpractice and plagiarism. *Clin Med* 2023 May;23(3):280 [FREE Full text] [doi: [10.7861/clinmed.Let.23.3.2](https://doi.org/10.7861/clinmed.Let.23.3.2)] [Medline: [37236804](https://pubmed.ncbi.nlm.nih.gov/37236804/)]
16. Liu J, Wang C, Liu S. Utility of ChatGPT in clinical practice. *J Med Internet Res* 2023 Jun 28;25:e48568 [FREE Full text] [doi: [10.2196/48568](https://doi.org/10.2196/48568)] [Medline: [37379067](https://pubmed.ncbi.nlm.nih.gov/37379067/)]
17. Ray PP. ChatGPT: a comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. *Internet of Things Cyber-Physical Syst* 2023;3:121-154. [doi: [10.1016/j.iotcps.2023.04.003](https://doi.org/10.1016/j.iotcps.2023.04.003)]
18. Milano S, McGrane JA, Leonelli S. Large language models challenge the future of higher education. *Nat Mach Intell* 2023 Mar 31;5(4):333-334. [doi: [10.1038/s42256-023-00644-2](https://doi.org/10.1038/s42256-023-00644-2)]
19. Chen J, Ran X. Deep learning with edge computing: a review. *Proc IEEE* 2019 Aug;107(8):1655-1674. [doi: [10.1109/jproc.2019.2921977](https://doi.org/10.1109/jproc.2019.2921977)]
20. Bharany S, Sharma S, Khalaf OI, Abdulsahib GM, Al Humaimeedy AS, Aldhyani THH, et al. A systematic survey on energy-efficient techniques in sustainable cloud computing. *Sustainability* 2022 May 20;14(10):6256. [doi: [10.3390/su14106256](https://doi.org/10.3390/su14106256)]

21. Johnson SG, Potrebny T, Larun L, Ciliska D, Olsen NR. Usability methods and attributes reported in usability studies of mobile apps for health care education: scoping review. *JMIR Med Educ* 2022 Jun 29;8(2):e38259 [FREE Full text] [doi: [10.2196/38259](https://doi.org/10.2196/38259)] [Medline: [35767323](https://pubmed.ncbi.nlm.nih.gov/35767323/)]
22. Lu J, Schmidt M, Lee M, Huang R. Usability research in educational technology: a state-of-the-art systematic review. *Education Tech Research Dev* 2022 Aug 22;70(6):1951-1992. [doi: [10.1007/s11423-022-10152-6](https://doi.org/10.1007/s11423-022-10152-6)]
23. Hein HJ, Glombiewski JA, Rief W, Riecke J. Effects of a video intervention on physicians' acceptance of pain apps: a randomised controlled trial. *BMJ Open* 2022 Apr 25;12(4):e060020 [FREE Full text] [doi: [10.1136/bmjopen-2021-060020](https://doi.org/10.1136/bmjopen-2021-060020)] [Medline: [35470200](https://pubmed.ncbi.nlm.nih.gov/35470200/)]
24. Skalidis I, Muller O, Fournier S. CardioVerse: the cardiovascular medicine in the era of Metaverse. *Trends Cardiovasc Med* 2023 Nov;33(8):471-476 [FREE Full text] [doi: [10.1016/j.tcm.2022.05.004](https://doi.org/10.1016/j.tcm.2022.05.004)] [Medline: [35568263](https://pubmed.ncbi.nlm.nih.gov/35568263/)]
25. Lund BD, Wang T. Chatting about ChatGPT: how may AI and GPT impact academia and libraries? *Library High Tech News* 2023 Feb 14;40(3):26-29. [doi: [10.1108/lhtn-01-2023-0009](https://doi.org/10.1108/lhtn-01-2023-0009)]
26. Ma H, Zhang C, Bian Y, Liu L, Zhang Z, Zhao P, et al. Fairness-guided few-shot prompting for large language models. *arXiv*. 2023 Mar. URL: <https://arxiv.org/abs/2303.13217> [accessed 2023-12-20]
27. Hébert PC, MacDonald N, Flegel K, Stanbrook MB. Competing interests and undergraduate medical education: time for transparency. *CMAJ* 2010 Sep 07;182(12):1279-1279 [FREE Full text] [doi: [10.1503/cmaj.100605](https://doi.org/10.1503/cmaj.100605)] [Medline: [20457768](https://pubmed.ncbi.nlm.nih.gov/20457768/)]
28. Wu Z, Merrill W, Peng H, Beltagy I, Smith NA. Transparency helps reveal when language models learn meaning. *Trans Assoc Comput Ling* 2023;11:617-634 [FREE Full text] [doi: [10.1162/tacl_a_00565](https://doi.org/10.1162/tacl_a_00565)]
29. Susnjak T. Beyond predictive learning analytics modelling and onto explainable artificial intelligence with prescriptive analytics and ChatGPT. *Int J Artif Intell Educ* 2023 Jun 22:1-31 [FREE Full text] [doi: [10.1007/s40593-023-00336-3](https://doi.org/10.1007/s40593-023-00336-3)]
30. Du M, He F, Zou N, Tao D, Hu X. Shortcut learning of large language models in natural language understanding. *Commun ACM* 2023 Dec 21;67(1):110-120. [doi: [10.1145/3596490](https://doi.org/10.1145/3596490)]
31. Mishra V, Singh J. Health technology assessment of telemedicine interventions in diabetes management: evidence from UAE. *FIIB Bus Rev* 2022 Nov 29;231971452211306. [doi: [10.1177/23197145221130651](https://doi.org/10.1177/23197145221130651)]
32. Dua S, Sharma MG, Mishra V, Kulkarni SD. Modelling perceived risk in blockchain enabled supply chain utilizing fuzzy-AHP. *J Glob Oper Strateg Sourc* 2022 Aug 10;16(1):161-177. [doi: [10.1108/jgoss-06-2021-0046](https://doi.org/10.1108/jgoss-06-2021-0046)]
33. Mishra V, Rana S. Understanding barriers to inbound medical tourism in the United Arab Emirates from a provider's perspective. *Worldw Hosp Tour Themes* 2022 Nov 30;15(2):131-142. [doi: [10.1108/whatt-10-2022-0122](https://doi.org/10.1108/whatt-10-2022-0122)]
34. Ahmed F, Mishra V. Estimating relative immediacy of water-related challenges in Small Island Developing States (SIDS) of the Pacific Ocean using AHP modeling. *Model Earth Syst Environ* 2019 Nov 02;6(1):201-214. [doi: [10.1007/s40808-019-00671-2](https://doi.org/10.1007/s40808-019-00671-2)]
35. Groenland E. *Qualitative methodologies and data collection methods: Toward increased rigour in management research*. Singapore: World Scientific; 2019.
36. Sushil. Interpreting the Interpretive Structural Model. *Glob J Flex Syst Manag* 2012 Sep 18;13(2):87-106. [doi: [10.1007/s40171-012-0008-3](https://doi.org/10.1007/s40171-012-0008-3)]
37. Prasad UC, Suri RK. Modeling of continuity and change forces in private higher technical education using total interpretive structural modeling (TISM). *Global J Flexible Syst Manage* 2017 Oct 4;12(3-4):31-39. [doi: [10.1007/bf03396605](https://doi.org/10.1007/bf03396605)]
38. Cacciamani G, Eppler MB, Ganjavi C, Pekan A, Biedermann B, Collins GS, et al. Development of the ChatGPT, generative artificial intelligence and natural large language models for accountable reporting and use (CANGARU) guidelines. *arXiv*. 2023 Jul. URL: <https://arxiv.org/abs/2307.08974> [accessed 2023-12-20]
39. Jamal A, Solaiman M, Alhasan K, Temsah MH, Sayed G. Integrating ChatGPT in medical education: adapting curricula to cultivate competent physicians for the AI era. *Cureus* 2023 Aug;15(8):e43036 [FREE Full text] [doi: [10.7759/cureus.43036](https://doi.org/10.7759/cureus.43036)] [Medline: [37674966](https://pubmed.ncbi.nlm.nih.gov/37674966/)]
40. Tan LF, Heng JJY, Teo DB. Response to: "The next paradigm shift? ChatGPT, artificial intelligence, and medical education". *Medical Teacher* 2023 Sep 13;46(1):151-152. [doi: [10.1080/0142159x.2023.2256961](https://doi.org/10.1080/0142159x.2023.2256961)]
41. Rudolph J, Tan S, Tan S. War of the chatbots: Bard, Bing Chat, ChatGPT, Ernie and beyond. The new AI gold rush and its impact on higher education. *J Appl Learn Teach* 2023 Apr 25;6(1):364-389 [FREE Full text] [doi: [10.37074/jalt.2023.6.1.23](https://doi.org/10.37074/jalt.2023.6.1.23)]
42. Baslom MMM, Tong S. Strategic management of organizational knowledge and employee's awareness about artificial intelligence with mediating effect of learning climate. *Int J Comput Intell Syst* 2019;12(2):1585. [doi: [10.2991/ijcis.d.191025.002](https://doi.org/10.2991/ijcis.d.191025.002)]
43. Rao A, Pang M, Kim J, Kamineni M, Lie W, Prasad AK, et al. Assessing the utility of ChatGPT throughout the entire clinical workflow: development and usability study. *J Med Internet Res* 2023 Aug 22;25:e48659 [FREE Full text] [doi: [10.2196/48659](https://doi.org/10.2196/48659)] [Medline: [37606976](https://pubmed.ncbi.nlm.nih.gov/37606976/)]
44. Zhang J, Krishna R, Awadallah AH, Wang C. EcoAssistant: using LLM Assistant more affordably and accurately. *arXiv*. 2023. URL: <https://arxiv.org/abs/2310.03046> [accessed 2023-12-20]
45. Memarian B, Doleck T. Fairness, Accountability, Transparency, and Ethics (FATE) in artificial intelligence (AI) and higher education: a systematic review. *Comput Educ Artif Intell* 2023;5:100152. [doi: [10.1016/j.caeai.2023.100152](https://doi.org/10.1016/j.caeai.2023.100152)]

46. Mishra V. Data for AHP and TISM models for the CUC-FATE framework. GitHub. URL: https://github.com/vinaytosh/datasharing/blob/master/Data_CUCFATE.xlsx [accessed 2023-12-20]

Abbreviations

AHP: analytical hierarchy process

AI: artificial intelligence

CI: consistency index

CR: consistency ratio

CUC-FATE: Cost, Usability, Credibility, Fairness, Accountability, Transparency, and Explainability

FATE: Fairness, Accountability, Transparency, and Explainability

LLM: large language model

MICMAC: cross-impact matrix multiplication applied to classification

NLP: natural language processing

RI: random index

TISM: total interpretive structural modeling

Edited by K El Emam; submitted 16.08.23; peer-reviewed by S Sedaghat, B Senst, M Pandey, S Kulkarni; comments to author 11.12.23; revised version received 20.12.23; accepted 03.02.24; published 23.04.24.

Please cite as:

Quttainah M, Mishra V, Madakam S, Lurie Y, Mark S

Cost, Usability, Credibility, Fairness, Accountability, Transparency, and Explainability Framework for Safe and Effective Large Language Models in Medical Education: Narrative Review and Qualitative Study

JMIR AI 2024;3:e51834

URL: <https://ai.jmir.org/2024/1/e51834>

doi: [10.2196/51834](https://doi.org/10.2196/51834)

PMID: [38875562](https://pubmed.ncbi.nlm.nih.gov/38875562/)

©Majdi Quttainah, Vinaytosh Mishra, Somayya Madakam, Yotam Lurie, Shlomo Mark. Originally published in JMIR AI (<https://ai.jmir.org>), 23.04.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR AI, is properly cited. The complete bibliographic information, a link to the original publication on <https://www.ai.jmir.org/>, as well as this copyright and license information must be included.

Review

Approaches for the Use of AI in Workplace Health Promotion and Prevention: Systematic Scoping Review

Martin Lange¹, Prof Dr; Alexandra Löwe¹, MA; Ina Kayser², Prof Dr; Andrea Schaller³, Prof Dr

¹Department of Fitness & Health, IST University of Applied Sciences, Duesseldorf, Germany

²Department of Communication & Business, IST University of Applied Sciences, Duesseldorf, Germany

³Institute of Sport Science, Department of Human Sciences, University of the Bundeswehr Munich, Munich, Germany

Corresponding Author:

Martin Lange, Prof Dr

Department of Fitness & Health

IST University of Applied Sciences

Erkrather Straße 220a-c

Duesseldorf, 40233

Germany

Phone: 49 211 86668 ext 656

Email: mlange@ist-hochschule.de

Abstract

Background: Artificial intelligence (AI) is an umbrella term for various algorithms and rapidly emerging technologies with huge potential for workplace health promotion and prevention (WHPP). WHPP interventions aim to improve people's health and well-being through behavioral and organizational measures or by minimizing the burden of workplace-related diseases and associated risk factors. While AI has been the focus of research in other health-related fields, such as public health or biomedicine, the transition of AI into WHPP research has yet to be systematically investigated.

Objective: The systematic scoping review aims to comprehensively assess an overview of the current use of AI in WHPP. The results will be then used to point to future research directions. The following research questions were derived: (1) What are the study characteristics of studies on AI algorithms and technologies in the context of WHPP? (2) What specific WHPP fields (prevention, behavioral, and organizational approaches) were addressed by the AI algorithms and technologies? (3) What kind of interventions lead to which outcomes?

Methods: A systematic scoping literature review (PRISMA-ScR [Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews]) was conducted in the 3 academic databases PubMed, Institute of Electrical and Electronics Engineers, and Association for Computing Machinery in July 2023, searching for papers published between January 2000 and December 2023. Studies needed to be (1) peer-reviewed, (2) written in English, and (3) focused on any AI-based algorithm or technology that (4) were conducted in the context of WHPP or (5) an associated field. Information on study design, AI algorithms and technologies, WHPP fields, and the patient or population, intervention, comparison, and outcomes framework were extracted blindly with Rayyan and summarized.

Results: A total of 10 studies were included. Risk prevention and modeling were the most identified WHPP fields (n=6), followed by behavioral health promotion (n=4) and organizational health promotion (n=1). Further, 4 studies focused on mental health. Most AI algorithms were machine learning-based, and 3 studies used combined deep learning algorithms. AI algorithms and technologies were primarily implemented in smartphone apps (eg, in the form of a chatbot) or used the smartphone as a data source (eg, Global Positioning System). Behavioral approaches ranged from 8 to 12 weeks and were compared to control groups. Additionally, 3 studies evaluated the robustness and accuracy of an AI model or framework.

Conclusions: Although AI has caught increasing attention in health-related research, the review reveals that AI in WHPP is marginally investigated. Our results indicate that AI is promising for individualization and risk prediction in WHPP, but current research does not cover the scope of WHPP. Beyond that, future research will profit from an extended range of research in all fields of WHPP, longitudinal data, and reporting guidelines.

Trial Registration: OSF Registries osf.io/bfswp; <https://osf.io/bfswp>

(JMIR AI 2024;3:e53506) doi:[10.2196/53506](https://doi.org/10.2196/53506)

KEYWORDS

artificial intelligence; AI; machine learning; deep learning; workplace health promotion; prevention; workplace health promotion and prevention; technology; technologies; well-being; behavioral health; workplace-related; public health; biomedicine; PRISMA-ScR; Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews; WHPP; risk; AI-algorithm; control group; accuracy; health-related; prototype; systematic review; scoping review; reviews; mobile phone

Introduction

Artificial Intelligence as an Umbrella Concept

Artificial intelligence (AI) is a concept that dates back to the mid-1900s [1] and was first defined as “the science and engineering of making intelligent machines” [2]. Today, AI is described as a computer system’s capability to perform complex tasks that mimic human cognitive functions to perform tasks such as reasoning, decision-making, or problem-solving, autonomously and adaptively [3]. However, its capabilities and underlying functions have changed significantly over the decades [1,4]. More recently, AI has emerged as a transformative force across various industries. Its application has shown promise in health promotion and health care [5-7], opening new possibilities concerning patient care and enhanced medical practices.

There is growing consensus in the literature that adaptivity and autonomy are the key characteristics of AI applications and technologies [5]. AI is considered an umbrella concept of emerging technologies, enclosing fundamental distinct types such as machine learning (ML), deep learning (DL), or natural language processing (NLP) [4,8]. Technically, AI is an ML-based approach that simulates human minds’ cognitive and affective functions [3,8] and is designed to observe and react to a specific environment. In contrast to deterministic programming, such models feature many free parameters that can adapt autonomously to calibrate the model. For example, AI can be applied in repetitive tasks requiring human intelligence, such as scanning and interpreting magnetic resonance imaging, autonomous driving, or analyzing big data sets [9-11]. ML and DL algorithms and artificial neural networks enable a machine or system to learn from large data sets, make autonomous decisions, and improve their performance over time [4]. More narrowly, NLP allows machines to generate and understand text and spoken language in the same way humans do. It combines rule-based natural language modeling with ML and DL models to process human language in text or speech data, understand its meaning, including feelings, and even generate human language, as it is sometimes used in chatbots or language translation [12].

AI in Health Care and Public Health

Implementing AI algorithms and technologies for health care institutions bears enormous potential, ranging from efficient health service management, predictive medicine, patient data, and diagnostics with real-time analyses to clinical decision-making. Most studies report a broader AI architecture with a combination of algorithms rooted in ML, DL, and NLP [4,11]. For example, 1 AI approach evaluated the support of clinical decision-making by analyzing continuous laboratory data, past clinical notes, and current information of physicians synthesizing significant associations [13]. AI implementation

in the form of predictive modeling showed positive results by detecting irregular heartbeats through smartwatches [14], automatically identifying reports of infectious disease in the media [15], or ascertaining cardiovascular risk factors from retinal images [16]. Through systematic profiling of 4518 existing drugs against 578 cancer cell lines with an AI-based approach, a study revealed that nononcology drugs have an unexpectedly high rate of anticancer activity [17]. Another study developed and evaluated a Medical Instructed Real-Time Assistant that listens to the user’s chief complaint and predicts a specific disease [18]. Chatbots have been used to detect COVID-19 symptoms through detailed questioning [6] or to predict the risk of type II diabetes mellitus [19].

Workplace Health Promotion and Prevention

As adults spend a significant amount of time working, it is widely accepted that work and work environments have a major impact on individuals’ health. Workplace health promotion and prevention (WHPP) are important fields that “[...] improve the health and well-being of people at work [...]” [20] through a combination of behavioral and organizational measures. Workplace health promotion follows a competence-oriented, salutogenetic approach to promoting the resources of an individual [20]. Prevention in the workplace focuses on minimizing the burden of workplace-related diseases and associated risk factors [21,22]. WHPP interventions range from behavioral measures with active participation (eg, courses or seminars) to organizational measures such as consultations, analyses, inspections, and establishing organizational structures such as a health committee [23,24].

Prior Work

With the Luxembourg declaration, WHPP has evolved into an independent discipline that differentiates from return-to-work (RTW) and occupational safety and health (OSH) measures [20,25]. In OSH-related disciplines, previous reviews have focused on risk assessment or detection related to physical ergonomics [26], occupational physical fatigue [27], or core body temperature [28]. Other reviews explored the evidence of AI in F-related areas, such as vocational rehabilitation [29] and functional capacity evaluation [30]. In health promotion in general, 1 review evaluates the use of chatbots to increase health-related behavior but does not focus on the workplace setting [31]. To the authors’ knowledge, no review has evaluated the use of AI in WHPP.

Therefore, this systematic scoping review aims to comprehensively assess an overview of the current use of AI in WHPP. The results will then be used to point to future research directions. The following research questions (RQ) were derived from these aims:

- RQ1: What are the study characteristics of studies on AI algorithms and technologies in WHPP?

- RQ2: What specific WHPP fields (prevention, behavioral, and organizational approaches) are addressed by the AI algorithms and technologies?
- RQ3: What kind of interventions were conducted, and what outcomes were assessed?

Methods

Design

A systematic scoping review approach [32] was selected following the extended PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews; [Multimedia Appendix 1](#)) [33]. We applied the 5-step framework to identify current or emerging research directions and provide an overview of research activities [34]. Additionally, the patient or population, intervention, comparison, and outcomes (PICO) framework [35] was used to specify the study's objective, from the search string and data charting to more systematic discussion [36]. The review was registered prospectively in the Open Science Framework (OSF) on July 5, 2023. All files (protocol, search string, and search results) have been uploaded to the OSF profile and are publicly accessible [37].

Eligibility Criteria

Included studies needed to be (1) peer-reviewed, (2) written in English, and (3) focused on any AI-based algorithm or technology that (4) were conducted in the context of WHPP, or (5) an associated field (workplace prevention, occupational health, and workplace health) that applies to WHPP. The types of research considered were review types (systematic, scoping, or rapid), cross-sectional studies, and longitudinal studies.

Our conceptualization of AI included the concepts of “machine learning,” “deep learning,” and “natural language processing.” Our conceptualization of “workplace health promotion and prevention” followed a broader understanding comprising the setting (eg, “work,” “workplace,” or “in or at the workplace”), the target population (eg, “working adults” or “employees”) and the outcome dimension (eg, “health” or “health behavior”). The search period was limited to studies published since January 2000 and before July 31, 2023. During the review, the search was extended to December 20, 2023.

Information Sources and Search

The systematic literature research was conducted in July 2023 in 3 databases: PubMed, IEEE Xplore, and Association for

Computing Machinery. The search string included Boolean operators (“AND,” “OR,” and “NOT”) and search terms related to “artificial intelligence,” “workplace health promotion,” “health promotion,” and “workplace setting” (see supplementary files available at OSF profile [37]). Papers were managed with the software tool Rayyan, followed by a 2-stage screening process. First, 1 reviewer (ML) removed all duplicates. Second, 2 reviewers (ML and AL) screened all titles or abstracts and read full texts for eligibility criteria in a blinded procedure. Disagreement was resolved by either consensus of the 2 reviewers or by consultation of a third reviewer (IK).

Data Charting and Synthesis of Results

In the first step, the study characteristics were extracted: first author (name and year), study design (eg, cross-sectional or randomized controlled trial), the primary type of AI algorithm and technology as referred to in the study (eg, AI, ML, DL, or NLP), and the frontend in which the AI-technology was implemented (eg, mobile app or web app). Second, the PICO framework [35] was applied to extract information about the target group (number of included participants/workplace context), the intervention approach, the comparison, and the reported outcomes of the study.

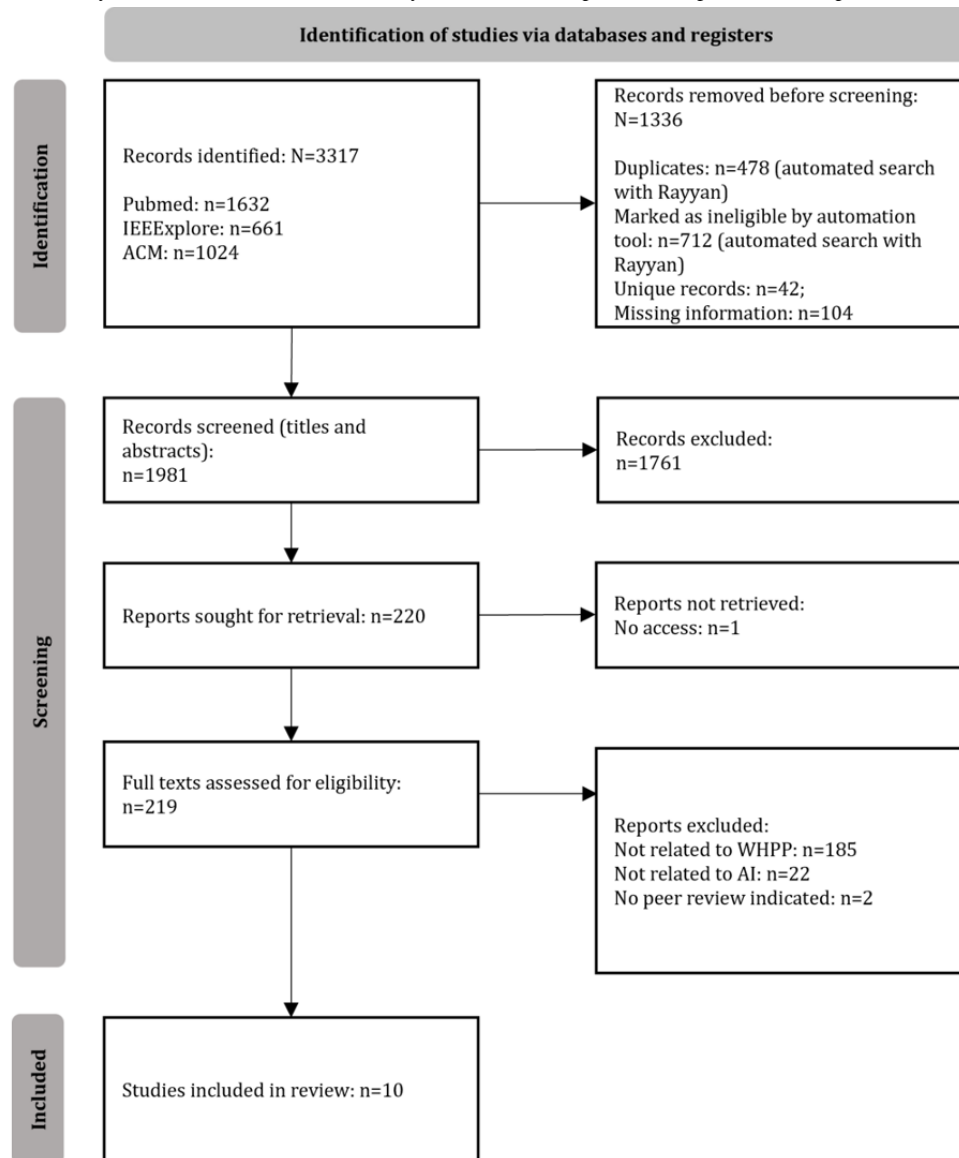
We used the extracted information from the study characteristics to answer RQ1 on current AI-based technologies applied in WHPP. For answering RQ2 and RQ3, we used the data extracted by the PICO framework. The information was then categorized within the results' tables and summarized narratively.

Results

Included Studies

The predefined search led to a total of 3317 results. The screening results revealed 478 duplicates, 712 records not meeting inclusion criteria (eg, publication type, language, or setting), 42 unique records, and 104 with missing information, leaving 1981 records for the title and abstract screening. The title and abstract screening excluded another 1761 records for not meeting inclusion criteria, leading to 220 records for full-text screening, of which one was inaccessible. After screening 219 full-text records, another 209 records were excluded. Finally, 10 studies remained in this systematic scoping review (the PRISMA-ScR flowchart is shown in [Figure 1](#)).

Figure 1. PRISMA flowchart of the literature search process. ACM: Association for Computing Machinery; AI: artificial intelligence; PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses; WHPP: workplace health promotion and prevention.



Study Characteristics (RQ1)

The results of the study characteristics are presented in [Table 1](#). Regarding the study designs, 6 studies were cross-sectional studies [38-43], 3 were randomized controlled trials [44-46], and 1 was a quasi-controlled trial [47]. None of the studies explained data protection standards (security protocols, storage location or duration, or access of third parties) within the AI algorithms and technologies used. In most studies, white-collar workers were the intended target group [38,41,42,46], whereas, in 3 studies, white-collar and physical labor workers participated [40,45,47]. Further, 1 study evaluated AI-based technologies with physical labor workers [39], and another did not disclose

any information about the type of work setting [44]. Information on sample characteristics was missing in 3 studies [40,41,44], little information was provided in 2 studies [38,44], and 4 studies offered sufficient information [39,42].

A comparison was used in different ways by 6 studies [40,42,44-47]. Further, 4 studies recruited a classic control group [39,44,46,47], 2 of which exposed the control group after a waiting period [44,46]. Another study compared their assessed data to external data thresholds [40], and 1 study compared assessed objective data with subjective data [42]. Regarding the outcome, all studies stated sufficient and significant results. Further, 1 study reported no changes in 1 of the 3 assessed outcomes [47].

Table 1. Study characteristics, AI^a algorithms and technologies, and WHPP^b fields.

Author	Year	Included type of AI algorithm	Implemented frontend	WHPP field	Study design
Anan et al [45]	2021	Machine learning	Smartphone app with integrated chatbot	Prevention; behavioral health promotion	RCT ^c
Morshed et al [38]	2022	Machine learning	Software-based sensor technology	Prevention	CS ^d
Cui et al [39]	2020	Deep learning networks (recurrent neural network or long-short-term neural network)	N/A ^e	Prevention (risk assessment)	CS
Dijkhuis et al [44]	2018	Machine learning	Web app	Behavioral health promotion	RCT
Hungerbuehler et al [40]	2021	Machine learning	Viki chatbot within a web browser interface	Prevention (risk assessment)	CS
Kaiser et al [41]	2021	Fuzzy neural network-based fusion	Smartphone app with GPS ^f and eHealth sensor	Organizational health promotion (risk assessment)	CS
Lopes et al [47]	2023	Neural language processing or machine learning	EMYS ^g robot	Behavioral health promotion	qCT ^h
Maxhuni et al [42]	2021	Machine learning	Smartphone app	Prevention (risk assessment)	CS
Piao et al [46]	2020	Deep learning networks, machine learning, and natural language processing (large language model)	Watson conversation tool (IBM Corp) integrated into a smartphone app	Behavioral health promotion	RCT
Yan et al [43]	2020	Convolutional neural network	Web-based app	Prevention (risk assessment)	CS

^aAI: artificial intelligence.

^bWHPP: workplace health promotion and prevention.

^cRCT: randomized controlled trial.

^dCS: cross-sectional study design.

^eN/A: not applicable.

^fGPS: Global Positioning System.

^gEMYS: emotive head system.

^hqCT: quasi controlled trial.

AI Applications and Technologies in Specific WHPP Fields (RQ2)

AI algorithms and technologies were mainly used for preventive purposes in risk assessment (Table 1). Furthermore, 2 studies evaluated prediction models [39,42]. Additionally, 3 studies [44,46,47] targeted health behavior change using 3 different approaches ranging from a web app [44] and smartphone app [46] to social robot agents [47]. Further, 1 study [41] was categorized as an organizational health promotion approach. A major target indication was mental health, which was addressed in 4 studies [38,40,42,43]. In contrast, 1 study dealt with musculoskeletal disorders [45] and 1 on overall physical health and work-related factors [39].

Interventions and Outcomes (RQ3)

The PICO category “intervention” did not apply to studies focusing on prevention since they did not evaluate an intervention [38-43]. Interventions were evaluated by 4 studies [44-47] with a duration of 12 weeks [44-46] and 8 weeks [47]. Within these 4 studies, 2 used chatbots as a primary AI application [45,46], 1 used a web application [44], and 1 used a social robot agent [47]. These 4 studies recruited a control group, of which 2 studies exposed the control group after a waiting period [44,46]. Regarding the outcome, all studies stated sufficient and significant results. The study of Lopes et al [47] reported no changes in 1 of the 3 assessed outcomes (Table 2).

Table 2. Interventions and outcomes of studies included in the review.

	Population	Intervention	Comparison	Outcome
Anan et al [45]	IG ^a 48 and CG ^b 46 engineers and white-collar workers	AI ^c -assisted program for MSD ^d that selects exercises depending on participants' chat input; 12-week intervention with individualized exercises for stretching, maintaining good posture, and mindfulness.	CG: exercise routine of 3 minutes per day during break time; routine consists of standard exercises for stretching, maintaining good posture, and mindfulness.	Adherence rate: 92%; significant difference in the worst pain scores of neck or shoulder pain or stiffness and low back pain between baseline and 12 weeks (score: -1.12; 95% CI -1.53 to -0.70; $P < .001$); significant improvements of IG in the severity of the neck or shoulder pain or stiffness and low back pain compared to CG (OR ^e 6.36, 95% CI 2.57-15.73; $P < .001$); subjective improvement in symptoms in IG at 12 weeks (score: 43; 95% CI 11.25-164.28; $P < .001$).
Morshed et al [38]	46 remote information workers	Development and implementation of a workplace stress sensing system for 4 weeks using passive sensors (email, calendar, app, mouse and keyboard use; facial positions and facial action units; or physiological sensors).	Comparison of passive sensor data with self-report (study intake, experience sampling, daily check-in, daily check-out, end of study expectations) data.	Passive sensors detect triggers and manifestations of workplace stress effectively (eg, keyboard activity and less facial movement were positively correlated with stress ($r = -0.05$, $P < .05^f$ and $r = -0.09$, $P < .05^f$, respectively); the quality of stress models depends on prior data of the worker and the amount of data (F_1 -score: after 10 days=58%; after 19 days=73%).
Cui et al [39]	4000 steel workers	Development and comparison of 2 AI-based risk prediction models (LSTM ^g vs RNN ^h) that predict the influence of the work environment on employees' health.	N/A ⁱ	Based on sociodemographic data (age, income, education, or marital status), health-related data (BMI, smoking, drinking, or blood lipids [cholesterol or triglyceride]), and work-related factors (length of service, high-temperature exposure, shift work, or noise exposure) the prediction effect of LSTM is significantly better than that of traditional RNN, with an accuracy of more than 95% (F_1 -score).
Dijkhuis et al [44]	IG 24 and CG 24 population/setting not disclosed	Development and implementation of a prediction model that personalizes physical activity recommendations. Within a 12-week workplace health promotion intervention. The goals of the intervention were to increase physical activity during workdays by improving physical and mental health and several work-related variables.	CG: no participation in the 12-week WHP ^j -program.	Input variables "hours of the day" and "step count" were used in the evaluated model and reached an accuracy of 90% (mean accuracy=0.93; range=0.88-0.99; mean F_1 -score=0.90; range=0.87-0.94). Tree algorithms and tree-based ensemble algorithms performed exceedingly well. The individualized algorithms allow for predicting physical activity during the day and provide the possibility to intervene with personalized feedback.
Hungerbuehler et al [40]	77 industrial, logistic, and office workers	Development of a chatbot system and its implementation in a workplace setting to assess employees' mental health.	Participation rates were compared to face-to-face collection method rates.	The response rate was 64.2% (77/120). The majority scored in the mild range for anxiety (GAD-7 ^k : mean 6.21, SD 4.56; 50%) and depression (PHQ-9 ^l : mean 4.40, SD 5.21; 57%), the moderate range for stress (DASS-21 ^m : mean 11.09, SD 7.13; 46%), subthreshold level for insomnia (ISI ⁿ : mean 9.26, SD 5.66; 70%), the low-risk burnout-category (OLBI ^o : mean 27.68, SD 8.38; 68%) and in the increased risk category for stress (JSSP ^p : mean 32.38, SD 3.55; 69%). Chatbot-based workplace mental health assessment is highly engaging and effective among employees, with response rates comparable to face-to-face interviews.
Kaiser et al [41]	12 office workers	Evaluation of a portable health (pHealth) app to detect COVID-19 infection and trace movement to prevent further infections. Additionally, the pHealth app detects employees' health conditions and recommends further health measures if indicated.	N/A	The app-integrated COVID-19 questionnaire was validated against real-time health conditions. Proximity detection, contact tracing, and health monitoring (external sensors) were confirmed by proximity testing (surf plot evaluation); it effectively estimates COVID-19 infection risk and personal health conditions.

	Population	Intervention	Comparison	Outcome
Lopes et al [47]	IG 28 and CG 28 service and retail workers	IG interacted with a social robot agent that promotes health behavior change of participants' choice (physical activity, nutrition, tobacco consumption, and stress and anxiety) in the workplace. After baseline assessment 8, social robots were used for 20-30 minutes weekly for 8 weeks. Based on the health action process approach model, the intervention focused on goal setting, monitoring behavior, elaborating action plans, and self-efficacy techniques through videos.	CG received the same intervention measures through human agents via Teams (Microsoft Corp).	IG improved significantly compared to CG in productivity ($F_{1,46}=9041, P<.005^f, \eta^2=0.26$) and in well-being ($F_{1,53}=4517, P<.005^f, \eta^2=0.079$), but not in work-engagement ($F_{1,49}=0.5176, P>.005^f$). Additionally, IG improved significantly in the postintervention scores compared to CG ($F_{1,43}=8997, P<.001^f, \text{Wilk } \Lambda=0.597, \text{partial } \eta^2=0.40$) despite presenteeism and regard for their level of mental well-being.
Maxhuni et al [42]	30 office workers	Measurement of smartphone data to assess employees' stress levels. Data were assessed for 8 weeks on physical activity (accelerometer), location (GPS ^d), social interaction (microphone, number of phone calls, or text messages), and social activity (app usage).	Objective data was compared to subjective data (OLBI, POMS ^f).	A high correlation between objective smartphone data and questionnaire scores was overall significant. The accuracy of the supervised decision tree was acceptable ($F_1\text{-score}=67.5\%$). The semisupervised learning approach was somewhat better, with an $F_1\text{-score}$ of 70%. Overall, the results confirm that the prediction model is feasible to detect perceived stress at work using smartphone-sensed data.
Piao et al [46]	IG 57 and CG 49 office and administrative workers	A healthy lifestyle coaching chatbot from the KakaoTalk App (Kakao Corp) was implemented into an office work setting to promote employees' stair-climbing habits. During the intervention, the IG received cues, intrinsic, and extrinsic rewards for the entire 12 weeks.	CG did not receive intrinsic rewards for the first 4 weeks and only received all rewards, as in IG, from the fifth to the 12th week.	After 4 weeks, the change in SRHI ^s scores was (mean IG 13.54, SD 14.99; mean CG 6.42, SD 9.42) significantly different between groups ($P<.05^f$). Between the fifth and 12th week, the change in SRHI scores of the intervention and control groups was comparable (mean IG 12.08, SD 10.87; mean CG 15.88, SD 13.29; $P=.21$). Level of physical activity showed a significant difference between the groups after 12 weeks of intervention ($F_{1,11}=21.16; P=.045$). Intrinsic reward was significantly influencing habit formation.
Yan et al [43]	352 respiratory therapists in medical centers and regional hospitals	Building a model to develop a web-based application for classifying mental illness at the workplace. Data on emotional labor and psychological health was assessed for 4 weeks with the ELMH ^t .	N/A	Model structure with 8 domains was confirmed with exploratory factor analysis, and 4 types of mental health were classified using the Rasch analysis with an accuracy rate of $\text{MNSQ}^u=0.92$. An app predicting mental illness was successfully developed and demonstrated in this study.

^aIG: intervention group.

^bCG: control group.

^cAI: artificial intelligence.

^dMSD: musculoskeletal disorder.

^eOR: odds ratio.

^fOriginal P values were not reported in the original publications.

^gLSTM: long short-term memory.

^hRNN: recurrent neural network.

ⁱN/A: not applicable.

^jWHP: workplace health promotion.

^kGAD-7: Generalized Anxiety Disorder Scale.

^lPHQ-9: Physical Health Questionnaire.

^mDASS-21: Depression, Anxiety, Stress Scale.

ⁿISI: Insomnia Severity Index.

^oOLBI: Oldenburg Burnout Inventory.

^pJSS: job strain survey.

^qGPS: global positioning system.

^rPOMS: profile of mood states.

^sSRHI: self-report habit index.

^tELMH: Emotional Labor and Mental Health questionnaire.

^uMNSQ: mean square error.

Discussion

Principal Results

Overview

This study aimed to assess an overview of the current state of AI use in WHPP. Our results underline that despite the rapid increase in AI-related studies, only a small number of studies have addressed AI apps and technologies in WHPP up to now. Risk prediction and modeling were the most identified WHPP fields, followed by behavioral health promotion approaches. AI algorithms and technologies were primarily implemented in smartphone apps (eg, in the form of a chatbot) or used the smartphone as a data source (eg, GPS). Further, our results revealed that most studies validated AI algorithms and feasibility.

Potential Approaches

The results merely indicate the potential of AI in WHPP with individualized, real-time data analysis and health-related information as critical elements but do not fully reflect this at present. AI-assisted chatbot apps were a primary AI technology, reaching reasonable adherence rates and offering a potential access route through various frontend solutions such as smartphones or web-based apps. Chatbots can easily individualize health-related information and recommendations regarding the type of job, educational level, and specific language barriers. The integration of sensor technologies can increase the efficacy of individualized chatbot solutions. This could advance the access and dissemination of workplace health-related information significantly. Chronically ill employees or other target groups can profit from context-specific health information that helps maintain or improve workability [48]. The aspect of anonymity might increase the acceptance of prevention measures for smoking cessation, alcohol, or substance abuse [31,49]. Due to the diversity of job activities (eg, physical labor or white-collar jobs) and workplace characteristics (eg, office, hybrid, or remote work), individualized access to health interventions can improve resource allocation as well as the density and quality of preventive health care [50,51]. Personalizing health-related information or feedback potentially increases workplace health-related behaviors [52,53]. The genuine ability of AI to analyze large amounts of data in real-time can be applied to predict or detect individual or organizational health risks, for example, infections, stress symptoms, or body positions [54-59].

State of AI-Research in WHPP

The small number of studies on AI and WHPP compared to other sectors of work-related health (eg, OSH or RTW) or public health indicates a considerable research gap. At this point, research in other health care sectors offers much more reviews [7,60-62]. Reasons can be found in common challenges of WHPP as a young research field, a high sensitivity regarding data protection regulation in the context of work, and the nonexistent legal requirements for WHPP in many countries [23,63,64]. At the same time, WHPP is often entrenched within an OSH paradigm among employers that do not prioritize WHPP [65,66].

As stated, most research WHPP fields were prevention and risk prediction followed by behavioral approaches. Stress and mental health were the primary outcomes of 4 studies within these fields. Given the relevance of mental health, the research interest can be assessed as adequate. At the same time, musculoskeletal disorders are the leading cause of sick leave in most countries [67] and are therefore highly underrepresented in the included studies. In 2 studies, behavioral approaches focused on physical activity and general health behavior were investigated in 1 study. Other WHPP-related behaviors such as nutrition, sleep, substance abuse (eg, nicotine), or stress management are not targeted by current research [24]. The same accounts for organizational WHPP approaches centered in only 1 study [41]. Organizational approaches that aim to disseminate health-related information, increase work-related health literacy, or implement educational measures have not been included in current AI and WHPP research. Areas such as social inequality [68], specific target groups (eg, chronically ill employees or migrants), or health-oriented leadership were not addressed.

Most studies of our review were conducted in a cross-sectional study design to gain data for any AI learning process in a time- and resource-efficient way [69]. This has 2 implications regarding the current stage of research. First, AI model life cycles need to be completed to gain high-level semantics and create a comprehensive learning basis, from data preparation (eg, dealing with missing data) and data conditioning to data acquisition and model refinement [70]. For future AI models, longitudinal data are of utmost importance, as cross-sectional data can only reflect on a specific stage of that life cycle [70,71]. Second, longitudinal study designs are usually more cost- and resource-intensive and often less prioritized. This not only leads to an imbalance of evidence on behavioral WHPP interventions but also to a lack of causal relation between AI and WHPP outcomes.

Most studies reported using ML compared to more sophisticated DL or NLP algorithms. ML algorithms use extracted data to predict binary or multiple outcomes or classes without hidden layers. DL algorithms are characterized by hidden-layer neural networks. They can be employed for the analysis of more complex data sets, for example, for the detection of multidimensional objects in the realm of video and speech analysis [4,72]. The complexity of DL algorithms, in turn, ties in with the AI model life cycle, as DL algorithms require a broader database for learning. While ML approaches are found to be highly predictive and offer more individualized interventions in a specific context, they are also prone to errors. Escorpizio et al [29] point out that in 1 study, ML classification exceeded clinicians' decision-making [73]. Still, the results were later reversed when the approach was implemented with a different cohort [74]. This is of particular interest, as studies within our results relied on either a small number of participants [41], few input variables [44], or a homogenous data input (eg, only self-report data) [40], causing potential ceiling effects within the AI learning progress [75,76]. Conversely, the benefit of longitudinal data in the context of AI reveals itself through the increase in precision. Further, 1 study pointed out the relevance of multiple measurements and longitudinal data by increasing the accuracy from 46% (time point 0) to 73% after

19 days of data [38]. Nevertheless, the included studies do not use the potential of AI in comparable health-related fields such as OSH or RTW [26-31]. Some areas of AI application are not addressed, such as big data analysis (eg, comparison with existing data of national cohort studies) or language translation models.

Future Research

As pointed out, current research is on AI in WHPP regarding quantity, fields of WHPP and its subdomains, and AI algorithms. Future research should center around major causes of sick leave, such as musculoskeletal disorders, mental health, respiratory conditions, and influenza [67]. Behavioral WHPP interventions should extend to all areas of health-related behavior, including nutrition, sleep, substance abuse, and stress management [24]. Further, setting-specific aspects of WHPP, such as intervention content, implementation strategies, user experience, design, algorithms, and the company's size, need to be considered more specifically. So far, the studies have provided only moderate information on the job activities or the target groups. At the same time, workplaces and workers are diverse. The health of employees is influenced by numerous organizational and individual factors that must be further considered in the learning cycle of AI [77-79]. Regarding potential errors, existing AI algorithms must be validated with different target groups [59,80], emphasizing the need for longitudinal data and its impact on learning algorithms [81,82]. Beyond this, the technological diversity of the presented studies opens new possibilities for target group-specific or individualized interventions. Providing health information to chronically ill employees, migrants with different language skills, or individualizing health topics of varying age groups can be provided more effectively through AI to move beyond a "one size fits" all paradigm [83,84].

Outside of the objective's scope, we identified 2 aspects that can improve future research. First, the included studies reported overall positive results regarding feasibility, significance, or accuracy, underlining the vast potential that AI technology harbors. However, the results must be interpreted cautiously as certain information in the primary studies was not provided, assessed, or available at the stages of the investigated technology. For example, few studies mentioned a potential

bias through the novelty [40,47] or the Hawthorne effect [45,47,85]. The novelty effect [86] applies to most of the included studies as they did not control for experience with new technologies or their affinity to them. Second, concerns about data access, storage or control, the ownership of AI-generated data, and its further use need to be clarified [87,88]. Standards should be derived and updated at appropriate intervals, especially new AI-generated knowledge based on employee's personal information [89]. Transparency and high data protection regulation can increase adherence rates and reduce usage barriers [90]. In turn, we propose that future research should rely on reporting guidelines [76,91,92].

Strength and Limitations

Of note, 1 strength of our review is the explanatory nature of the RQs and the systematic search strategy in this new field. Consequently, the heterogeneity of the identified studies might be considered a limitation. Different AI applications and technologies, the types of intervention, and the variety of workplace settings limit the conclusion significantly. Beyond this, the reporting of the types of AI-based algorithms and technologies used in the study are based on the authors' self-reports. It is important to consider that the differentiation of the AI algorithm types cannot be made with a high degree of distinction.

Conclusions

Overall, this review underlines that AI in WHPP bears considerable potential but is not used fully at present. The results of our review offer a promising perspective on the predictive and personalized health paradigm shift in WHPP. Nevertheless, we conclude that current AI-related research in WHPP is still at the beginning, as it does not cover the scope of WHPP. The most salient research gaps can be found in lacking fields of WHPP and its subdomains, the predominantly ML-based algorithms and cross-sectional data, and the weak consideration of the work context. We believe we have contributed to future WHPP research by identifying these gaps and recommending future approaches. As AI applications are gaining an increasingly important role, we are convinced that future research will profit from an extended range of research in all fields of WHPP, longitudinal data, and the use of reporting guidelines.

Acknowledgments

The design and registration of the study was handled by ML. The first draft of this paper was by ML, AL, and AS. Data were collected by ML and AL. Analysis was done by ML, AL, and IK. Revision and review of this paper were performed by ML, AL, IK, and AS. This research received no external funding. We did not use any generative AI in this paper.

Data Availability

All data are publicly available in the OSF [37].

Conflicts of Interest

None declared.

Multimedia Appendix 1

PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) checklist. [PDF File (Adobe PDF File), 198 KB - [ai_v3i1e53506_app1.pdf](#)]

References

1. Kaul V, Enslin S, Gross SA. History of artificial intelligence in medicine. *Gastrointest Endosc* 2020 Oct;92(4):807-812. [doi: [10.1016/j.gie.2020.06.040](#)] [Medline: [32565184](#)]
2. McCarthy J. Programs with common sense mechanisation of thought processes. In: Proceedings of the Symposium of the National Physics Laboratory. London, UK: Her Majesty's Stationery Office; 1959 Presented at: Proceedings of the Symposium of the National Physics Laboratory; 24th-27th November 1958; Teddington, Middlesex p. 3-10.
3. Russell SJ, Norvig P. Introduction. In: Russell SJ, Norvig P, editors. *Artificial intelligence: a modern approach*. Harlow: Pearson; 2022:19-54.
4. Helm JM, Swiergosz AM, Haeberle HS, Karnuta JM, Schaffer JL, Krebs VE, et al. Machine learning and artificial intelligence: definitions, applications, and future directions. *Curr Rev Musculoskelet Med* 2020;13(1):69-76 [FREE Full text] [doi: [10.1007/s12178-020-09600-8](#)] [Medline: [31983042](#)]
5. Grossberg S. A path toward explainable AI and autonomous adaptive intelligence: deep learning, adaptive resonance, and models of perception, emotion, and action. *Front Neurobot* 2020;14:36 [FREE Full text] [doi: [10.3389/fnbot.2020.00036](#)] [Medline: [32670045](#)]
6. Chen J, See KC. Artificial intelligence for COVID-19: rapid review. *J Med Internet Res* 2020;22(10):e21476 [FREE Full text] [doi: [10.2196/21476](#)] [Medline: [32946413](#)]
7. Dong L, Yang Q, Zhang RH, Wei WB. Artificial intelligence for the detection of age-related macular degeneration in color fundus photographs: a systematic review and meta-analysis. *eClinicalMedicine* 2021;35:100875 [FREE Full text] [doi: [10.1016/j.eclinm.2021.100875](#)] [Medline: [34027334](#)]
8. Boucher P. Artificial intelligence: how does it work, why does it matter, and what can we do about it?. Brussels: European Parliament; 2020. URL: [https://www.europarl.europa.eu/RegData/etudes/STUD/2020/641547/EPRS_STU\(2020\)641547_EN.pdf](https://www.europarl.europa.eu/RegData/etudes/STUD/2020/641547/EPRS_STU(2020)641547_EN.pdf) [accessed 2024-07-30]
9. Lee S, Liu L, Radwin R, Li J. Machine learning in manufacturing ergonomics: recent advances, challenges, and opportunities. *IEEE Robot Autom Lett* 2021;6(3):5745-5752. [doi: [10.1109/ira.2021.3084881](#)]
10. Secinaro S, Calandra D, Secinaro A, Muthurangu V, Biancone P. The role of artificial intelligence in healthcare: a structured literature review. *BMC Med Inform Decis Mak* 2021;21(1):125 [FREE Full text] [doi: [10.1186/s12911-021-01488-9](#)] [Medline: [33836752](#)]
11. Johnson KB, Wei WQ, Weeraratne D, Frisse ME, Misulis K, Rhee K, et al. Precision medicine, AI, and the future of personalized health care. *Clin Transl Sci* 2021;14(1):86-93 [FREE Full text] [doi: [10.1111/cts.12884](#)] [Medline: [32961010](#)]
12. Raina V, Krishnamurthy S. *Building an effective data science practice*. Berkeley, CA: Apress; 2022.
13. Juhn Y, Liu H. Artificial intelligence approaches using natural language processing to advance EHR-based clinical research. *J Allergy Clin Immunol* 2020;145(2):463-469 [FREE Full text] [doi: [10.1016/j.jaci.2019.12.897](#)] [Medline: [31883846](#)]
14. Perez MV, Mahaffey KW, Hedlin H, Rumsfeld JS, Garcia A, Ferris T, Apple Heart Study Investigators. Large-scale assessment of a smartwatch to identify atrial fibrillation. *N Engl J Med* 2019;381(20):1909-1917 [FREE Full text] [doi: [10.1056/NEJMoa1901183](#)] [Medline: [31722151](#)]
15. Feldman J, Thomas-Bachli A, Forsyth J, Patel ZH, Khan K. Development of a global infectious disease activity database using natural language processing, machine learning, and human expertise. *J Am Med Inform Assoc* 2019;26(11):1355-1359 [FREE Full text] [doi: [10.1093/jamia/ocz112](#)] [Medline: [31361300](#)]
16. Poplin R, Varadarajan AV, Blumer K, Liu Y, McConnell MV, Corrado GS, et al. Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning. *Nat Biomed Eng* 2018;2(3):158-164. [doi: [10.1038/s41551-018-0195-0](#)] [Medline: [31015713](#)]
17. Corsello SM, Nagari RT, Spangler RD, Rossen J, Kocak M, Bryan JG, et al. Discovering the anti-cancer potential of non-oncology drugs by systematic viability profiling. *Nat Cancer* 2020;1(2):235-248 [FREE Full text] [doi: [10.1038/s43018-019-0018-6](#)] [Medline: [32613204](#)]
18. Rehman UU, Chang DJ, Jung Y, Akhtar U, Razzaq MA, Lee S. Medical instructed real-time assistant for patient with glaucoma and diabetic conditions. *Appl Sci* 2020;10(7):2216. [doi: [10.3390/app10072216](#)]
19. Jungwirth D, Haluza D. Artificial intelligence and public health: an exploratory study. *Int J Environ Res Public Health* 2023;20(5):4541 [FREE Full text] [doi: [10.3390/ijerph20054541](#)] [Medline: [36901550](#)]
20. Luxembourg declaration on workplace health promotion in the European Union. Perugia, Italy: European Network of Workplace Health Promotion; 2018.
21. Pomaki G, Franche RL, Murray E, Khushrushahi N, Lampinen TM. Workplace-based work disability prevention interventions for workers with common mental health conditions: a review of the literature. *J Occup Rehabil* 2012;22(2):182-195. [doi: [10.1007/s10926-011-9338-9](#)] [Medline: [22038297](#)]
22. Gritzka S, MacIntyre TE, Dörfel D, Baker-Blanc JL, Calogiuri G. The effects of workplace nature-based interventions on the mental health and well-being of employees: a systematic review. *Front Psychiatry* 2020;11:323 [FREE Full text] [doi: [10.3389/fpsy.2020.00323](#)] [Medline: [32411026](#)]

23. Terry PE. Workplace health promotion is growing up but confusion remains about what constitutes a comprehensive approach. *Am J Health Promot* 2019;33(6):845-849. [doi: [10.1177/0890117119854618](https://doi.org/10.1177/0890117119854618)] [Medline: [31159555](https://pubmed.ncbi.nlm.nih.gov/31159555/)]
24. Rongen A, Robroek SJW, van Lenthe FJ, Burdorf A. Workplace health promotion: a meta-analysis of effectiveness. *Am J Prev Med* 2013;44(4):406-415. [doi: [10.1016/j.amepre.2012.12.007](https://doi.org/10.1016/j.amepre.2012.12.007)] [Medline: [23498108](https://pubmed.ncbi.nlm.nih.gov/23498108/)]
25. Technical and ethical guidelines for workers' health surveillance. Geneva: International Labor Organization; 1998.
26. Donisi L, Cesarelli G, Pisani N, Ponsiglione AM, Ricciardi C, Capodaglio E. Wearable sensors and artificial intelligence for physical ergonomics: a systematic review of literature. *Diagnostics (Basel)* 2022;12(12):3048 [FREE Full text] [doi: [10.3390/diagnostics12123048](https://doi.org/10.3390/diagnostics12123048)] [Medline: [36553054](https://pubmed.ncbi.nlm.nih.gov/36553054/)]
27. Moshawrab M, Adda M, Bouzouane A, Ibrahim H, Raad A. Smart wearables for the detection of occupational physical fatigue: a literature review. *Sensors (Basel)* 2022;22(19):7472 [FREE Full text] [doi: [10.3390/s22197472](https://doi.org/10.3390/s22197472)] [Medline: [36236570](https://pubmed.ncbi.nlm.nih.gov/36236570/)]
28. Dolson CM, Harlow ER, Phelan DM, Gabbett TJ, Gaal B, McMellen C, et al. Wearable sensor technology to predict core body temperature: a systematic review. *Sensors (Basel)* 2022;22(19):7639 [FREE Full text] [doi: [10.3390/s22197639](https://doi.org/10.3390/s22197639)] [Medline: [36236737](https://pubmed.ncbi.nlm.nih.gov/36236737/)]
29. Escorpizo R, Theotokatos G, Tucker CA. A scoping review on the use of machine learning in return-to-work studies: strengths and weaknesses. *J Occup Rehabil* 2024;34(1):71-86. [doi: [10.1007/s10926-023-10127-1](https://doi.org/10.1007/s10926-023-10127-1)] [Medline: [37378718](https://pubmed.ncbi.nlm.nih.gov/37378718/)]
30. Fong J, Ocampo R, Gross DP, Tavakoli M. Intelligent robotics incorporating machine learning algorithms for improving functional capacity evaluation and occupational rehabilitation. *J Occup Rehabil* 2020;30(3):362-370. [doi: [10.1007/s10926-020-09888-w](https://doi.org/10.1007/s10926-020-09888-w)] [Medline: [32253595](https://pubmed.ncbi.nlm.nih.gov/32253595/)]
31. Aggarwal A, Tam CC, Wu D, Li X, Qiao S. Artificial intelligence-based chatbots for promoting health behavioral changes: systematic review. *J Med Internet Res* 2023;25:e40789 [FREE Full text] [doi: [10.2196/40789](https://doi.org/10.2196/40789)] [Medline: [36826990](https://pubmed.ncbi.nlm.nih.gov/36826990/)]
32. Peters MDJ, Godfrey CM, Khalil H, McInerney P, Parker D, Soares CB. Guidance for conducting systematic scoping reviews. *Int J Evid Based Healthc* 2015;13(3):141-146. [doi: [10.1097/XEB.0000000000000050](https://doi.org/10.1097/XEB.0000000000000050)] [Medline: [26134548](https://pubmed.ncbi.nlm.nih.gov/26134548/)]
33. Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, et al. PRISMA Extension for Scoping Reviews (PRISMA-ScR): checklist and explanation. *Ann Intern Med* 2018;169(7):467-473 [FREE Full text] [doi: [10.7326/M18-0850](https://doi.org/10.7326/M18-0850)] [Medline: [30178033](https://pubmed.ncbi.nlm.nih.gov/30178033/)]
34. Arksey H, O'Malley L. Scoping studies: towards a methodological framework. *Int J Soc Res Methodol* 2005;8(1):19-32. [doi: [10.1080/1364557032000119616](https://doi.org/10.1080/1364557032000119616)]
35. Huang X, Lin J, Demner-Fushman D. Evaluation of PICO as a knowledge representation for clinical questions. *AMIA Annu Symp Proc* 2006;2006:359-363 [FREE Full text] [Medline: [17238363](https://pubmed.ncbi.nlm.nih.gov/17238363/)]
36. Sager M, Pistone I. Mismatches in the production of a scoping review: highlighting the interplay of (in)formalities. *J Eval Clin Pract* 2019;25(6):930-937. [doi: [10.1111/jep.13251](https://doi.org/10.1111/jep.13251)] [Medline: [31368185](https://pubmed.ncbi.nlm.nih.gov/31368185/)]
37. Lange M, Löwe A, Kayser I, Schaller A. Approaches for the use of artificial intelligence in the field of workplace health: a systematic scoping review. *OSF*. 2023. URL: <https://osf.io/hsu2w/> [accessed 2023-10-06]
38. Morshed MB, Hernandez J, McDuff D, Suh J, Howe E, Rowan K, et al. Advancing the understanding and measurement of workplace stress in remote information workers from passive sensors and behavioral data. In: 10th International Conference on Affective Computing and Intelligent Interaction (ACII).: IEEE; 2022 Presented at: 2022 10th International Conference on Affective Computing and Intelligent Interaction (ACII); October 18-21, 2022; Nara, Japan p. 1-8. [doi: [10.1109/acii55700.2022.9953824](https://doi.org/10.1109/acii55700.2022.9953824)]
39. Cui S, Li C, Chen Z, Wang J, Yuan J. Research on risk prediction of dyslipidemia in steel workers based on recurrent neural network and LSTM neural network. *IEEE Access* 2020;8:34153-34161. [doi: [10.1109/access.2020.2974887](https://doi.org/10.1109/access.2020.2974887)]
40. Hungerbuehler I, Daley K, Cavanagh K, Garcia Claro H, Kapps M. Chatbot-based assessment of employees' mental health: design process and pilot implementation. *JMIR Form Res* 2021;5(4):e21678 [FREE Full text] [doi: [10.2196/21678](https://doi.org/10.2196/21678)] [Medline: [33881403](https://pubmed.ncbi.nlm.nih.gov/33881403/)]
41. Kaiser MS, Mahmud M, Noor MBT, Zenia NZ, Mamun SA, Mahmud KMA, et al. iWorksafe: towards healthy workplaces during COVID-19 with an intelligent phealth app for industrial settings. *IEEE Access* 2021;9:13814-13828. [doi: [10.1109/access.2021.3050193](https://doi.org/10.1109/access.2021.3050193)]
42. Maxhuni A, Hernandez-Leal P, Morales EF, Sucar LE, Osmani V, Mayora O. Unobtrusive stress assessment using smartphones. *IEEE Trans on Mobile Comput* 2021;20(6):2313-2325. [doi: [10.1109/tmc.2020.2974834](https://doi.org/10.1109/tmc.2020.2974834)]
43. Yan YH, Chien TW, Yeh YT, Chou W, Hsing SC. An app for classifying personal mental illness at workplace using fit statistics and convolutional neural networks: survey-based quantitative study. *JMIR mHealth uHealth* 2020;8(7):e17857 [FREE Full text] [doi: [10.2196/17857](https://doi.org/10.2196/17857)] [Medline: [32735232](https://pubmed.ncbi.nlm.nih.gov/32735232/)]
44. Dijkhuis TB, Blaauw FJ, van Ittersum MW, Velthuisen H, Aiello M. Personalized physical activity coaching: a machine learning approach. *Sensors (Basel)* 2018;18(2):623 [FREE Full text] [doi: [10.3390/s18020623](https://doi.org/10.3390/s18020623)] [Medline: [29463052](https://pubmed.ncbi.nlm.nih.gov/29463052/)]
45. Anan T, Kajiki S, Oka H, Fujii T, Kawamata K, Mori K, et al. Effects of an artificial intelligence-assisted health program on workers with neck/shoulder pain/stiffness and low back pain: randomized controlled trial. *JMIR mHealth uHealth* 2021;9(9):e27535 [FREE Full text] [doi: [10.2196/27535](https://doi.org/10.2196/27535)] [Medline: [34559054](https://pubmed.ncbi.nlm.nih.gov/34559054/)]

46. Piao M, Ryu H, Lee H, Kim J. Use of the healthy lifestyle coaching chatbot app to promote stair-climbing habits among office workers: exploratory randomized controlled trial. *JMIR mHealth uHealth* 2020;8(5):e15085 [FREE Full text] [doi: [10.2196/15085](https://doi.org/10.2196/15085)] [Medline: [32427114](https://pubmed.ncbi.nlm.nih.gov/32427114/)]
47. Lopes SL, Ferreira AI, Prada R. The use of robots in the workplace: conclusions from a health promoting intervention using social robots. *Int J Soc Robot* 2023;15:893-905 [FREE Full text] [doi: [10.1007/s12369-023-01000-5](https://doi.org/10.1007/s12369-023-01000-5)] [Medline: [37359429](https://pubmed.ncbi.nlm.nih.gov/37359429/)]
48. Schachner T, Keller R, V Wangenheim F. Artificial intelligence-based conversational agents for chronic conditions: systematic literature review. *J Med Internet Res* 2020;22(9):e20701 [FREE Full text] [doi: [10.2196/20701](https://doi.org/10.2196/20701)] [Medline: [32924957](https://pubmed.ncbi.nlm.nih.gov/32924957/)]
49. Ogilvie L, Prescott J, Carson J. The use of chatbots as supportive agents for people seeking help with substance use disorder: a systematic review. *Eur Addict Res* 2022;28(6):405-418 [FREE Full text] [doi: [10.1159/000525959](https://doi.org/10.1159/000525959)] [Medline: [36041418](https://pubmed.ncbi.nlm.nih.gov/36041418/)]
50. Xiao Z, Liao QV, Zhou M, Grandison T, Li Y. Powering an AI chatbot with expert sourcing to support credible health information access. In: *Proceedings of the 28th International Conference on Intelligent User Interfaces*. New York, NY, United States: Association for Computing Machinery; 2023 Presented at: 28th International Conference on Intelligent User Interfaces; 27th-31st March 2023; Sydney Australia p. 2-18 URL: <https://iui.acm.org/2023/> [doi: [10.1145/3581641.3584031](https://doi.org/10.1145/3581641.3584031)]
51. Jovanovic M, Baez M, Casati F. Chatbots as conversational healthcare services. *IEEE Internet Comput* 2021;25(3):44-51. [doi: [10.1109/mic.2020.3037151](https://doi.org/10.1109/mic.2020.3037151)]
52. Moore PV. OSH and the future of work: benefits and risks of artificial intelligence tools in workplaces. In: *Digital Human Modeling and Applications in Health, Safety, Ergonomics, and Risk Management : 10th International Conference, DHM 2019, Held as part of the 21st HCI International Conference, HCII 2019*. Orlando, FL, USA: Cham: Springer; 2019 Presented at: HCI International; 26th-31st July 2019; Orlando, Florida, United States of America p. 292-315 URL: <https://2019.hci.international/> [doi: [10.1007/978-3-030-22216-1_22](https://doi.org/10.1007/978-3-030-22216-1_22)]
53. Zhang J, Oh YJ, Lange P, Yu Z, Fukuoka Y. Artificial intelligence chatbot behavior change model for designing artificial intelligence chatbots to promote physical activity and a healthy diet: viewpoint. *J Med Internet Res* 2020;22(9):e22845 [FREE Full text] [doi: [10.2196/22845](https://doi.org/10.2196/22845)] [Medline: [32996892](https://pubmed.ncbi.nlm.nih.gov/32996892/)]
54. Conroy B, Silva I, Mehraei G, Damiano R, Gross B, Salvati E, et al. Real-time infection prediction with wearable physiological monitoring and AI to aid military workforce readiness during COVID-19. *Sci Rep* 2022;12(1):3797 [FREE Full text] [doi: [10.1038/s41598-022-07764-6](https://doi.org/10.1038/s41598-022-07764-6)] [Medline: [35260671](https://pubmed.ncbi.nlm.nih.gov/35260671/)]
55. Alberto R, Draicchio F, Varrecchia T, Silvetti A, Iavicoli S. Wearable monitoring devices for biomechanical risk assessment at work: current status and future challenges—a systematic review. *Int J Environ Res Public Health* 2018;15(9):2001 [FREE Full text] [doi: [10.3390/ijerph15092001](https://doi.org/10.3390/ijerph15092001)] [Medline: [30217079](https://pubmed.ncbi.nlm.nih.gov/30217079/)]
56. Saarela K, Huhta-Koivisto V, Kemell KK, Nurminen J. Work disability risk prediction using machine learning. In: Daimi K, Alsadoon A, Seabra Dos Reis S, editors. *Current and Future Trends in Health and Medical Informatics*. Cham: Springer Nature Switzerland; 2023:345-359.
57. Zawad MRS, Rony CSA, Haque MY, Banna MHA, Mahmud M, Kaiser MS. A hybrid approach for stress prediction from heart rate variability. In: *Frontiers of ICT in Healthcare: Proceedings of EAIT 2022*. Singapore: Springer Nature Singapore; 2023 Presented at: <https://www.csikolkata.org/eait2022/?i=1>; 30th-31st March 2022; Kolkata, India p. 111-121. [doi: [10.1007/978-981-19-5191-6_10](https://doi.org/10.1007/978-981-19-5191-6_10)]
58. Seo W, Kim N, Park C, Park SM. Deep learning approach for detecting work-related stress using multimodal signals. *IEEE Sensors J* 2022;22(12):11892-11902. [doi: [10.1109/jsen.2022.3170915](https://doi.org/10.1109/jsen.2022.3170915)]
59. Nijhawan T, Attigeri G, Ananthakrishna T. Stress detection using natural language processing and machine learning over social interactions. *J Big Data* 2022;9(1):33. [doi: [10.1186/s40537-022-00575-6](https://doi.org/10.1186/s40537-022-00575-6)]
60. Sarker S, Jamal L, Ahmed SF, Irtisam N. Robotics and artificial intelligence in healthcare during COVID-19 pandemic: a systematic review. *Rob Auton Syst* 2021;146:103902 [FREE Full text] [doi: [10.1016/j.robot.2021.103902](https://doi.org/10.1016/j.robot.2021.103902)] [Medline: [34629751](https://pubmed.ncbi.nlm.nih.gov/34629751/)]
61. Kumar Y, Koul A, Singla R, Ijaz MF. Artificial intelligence in disease diagnosis: a systematic literature review, synthesizing framework and future research agenda. *J Ambient Intell Humaniz Comput* 2023;14(7):8459-8486 [FREE Full text] [doi: [10.1007/s12652-021-03612-z](https://doi.org/10.1007/s12652-021-03612-z)] [Medline: [35039756](https://pubmed.ncbi.nlm.nih.gov/35039756/)]
62. Aggarwal R, Sounderajah V, Martin G, Ting DSW, Karthikesalingam A, King D, et al. Diagnostic accuracy of deep learning in medical imaging: a systematic review and meta-analysis. *NPJ Digit Med* 2021;4(1):65 [FREE Full text] [doi: [10.1038/s41746-021-00438-z](https://doi.org/10.1038/s41746-021-00438-z)] [Medline: [33828217](https://pubmed.ncbi.nlm.nih.gov/33828217/)]
63. Faller G. Future challenges for work-related health promotion in Europe: a data-based theoretical reflection. *Int J Environ Res Public Health* 2021;18(20):10996 [FREE Full text] [doi: [10.3390/ijerph182010996](https://doi.org/10.3390/ijerph182010996)] [Medline: [34682748](https://pubmed.ncbi.nlm.nih.gov/34682748/)]
64. Robroek SJ, Coenen P, Oude Hengel KM. Decades of workplace health promotion research: marginal gains or a bright future ahead. *Scand J Work Environ Health* 2021;47(8):561-564 [FREE Full text] [doi: [10.5271/sjweh.3995](https://doi.org/10.5271/sjweh.3995)] [Medline: [34655223](https://pubmed.ncbi.nlm.nih.gov/34655223/)]
65. Pescud M, Teal R, Shilton T, Slevin T, Ledger M, Waterworth P, et al. Employers' views on the promotion of workplace health and wellbeing: a qualitative study. *BMC Public Health* 2015;15:642 [FREE Full text] [doi: [10.1186/s12889-015-2029-2](https://doi.org/10.1186/s12889-015-2029-2)] [Medline: [26162910](https://pubmed.ncbi.nlm.nih.gov/26162910/)]

66. McCoy K, Stinson K, Scott K, Tenney L, Newman LS. Health promotion in small business: a systematic review of factors influencing adoption and effectiveness of worksite wellness programs. *J Occup Environ Med* 2014;56(6):579-587 [FREE Full text] [doi: [10.1097/JOM.0000000000000171](https://doi.org/10.1097/JOM.0000000000000171)] [Medline: [24905421](https://pubmed.ncbi.nlm.nih.gov/24905421/)]
67. Work-related MSDs: prevalence, costs and demographics in the EU. European Risk Observatory Executive summary. Luxembourg: European Agency for Safety and Health at Work (EU-OSHA); 2019. URL: https://osha.europa.eu/sites/default/files/Work_related_MSDs_prevalence_costs_and_demographics_in_EU_summary.pdf [accessed 2024-07-30]
68. van der Put AC, Mandemakers JJ, de Wit JBF, van der Lippe T. Worksite health promotion and social inequalities in health. *SSM Popul Health* 2020;10:100543 [FREE Full text] [doi: [10.1016/j.ssmph.2020.100543](https://doi.org/10.1016/j.ssmph.2020.100543)] [Medline: [32021901](https://pubmed.ncbi.nlm.nih.gov/32021901/)]
69. Wang X, Cheng Z. Cross-sectional studies: strengths, weaknesses, and recommendations. *Chest* 2020;158(1S):S65-S71. [doi: [10.1016/j.chest.2020.03.012](https://doi.org/10.1016/j.chest.2020.03.012)] [Medline: [32658654](https://pubmed.ncbi.nlm.nih.gov/32658654/)]
70. Ng MY, Kapur S, Blizinsky KD, Hernandez-Boussard T. The AI life cycle: a holistic approach to creating ethical AI for health decisions. *Nat Med* 2022;28(11):2247-2249 [FREE Full text] [doi: [10.1038/s41591-022-01993-y](https://doi.org/10.1038/s41591-022-01993-y)] [Medline: [36163298](https://pubmed.ncbi.nlm.nih.gov/36163298/)]
71. Challen R, Denny J, Pitt M, Gompels L, Edwards T, Tsaneva-Atanasova K. Artificial intelligence, bias and clinical safety. *BMJ Qual Saf* 2019;28(3):231-237 [FREE Full text] [doi: [10.1136/bmjqs-2018-008370](https://doi.org/10.1136/bmjqs-2018-008370)] [Medline: [30636200](https://pubmed.ncbi.nlm.nih.gov/30636200/)]
72. Lepakshi VA. Machine learning and deep learning based AI tools for development of diagnostic tools. In: Parihar A, Khan R, Kumar A, Kaushik A, Gohel H, editors. *Computational Approaches for Novel Therapeutic and Diagnostic Designing to Mitigate SARS-CoV-2 Infection*. Cambridge, Massachusetts, United States: Academic Press; 2022:399-420.
73. Gross DP, Zhang J, Steenstra I, Barnsley S, Haws C, Amell T, et al. Development of a computer-based clinical decision support tool for selecting appropriate rehabilitation interventions for injured workers. *J Occup Rehabil* 2013;23(4):597-609. [doi: [10.1007/s10926-013-9430-4](https://doi.org/10.1007/s10926-013-9430-4)] [Medline: [23468410](https://pubmed.ncbi.nlm.nih.gov/23468410/)]
74. Gross DP, Steenstra IA, Shaw W, Yousefi P, Bellinger C, Zaïane O. Validity of the work assessment triage tool for selecting rehabilitation interventions for workers' compensation claimants with musculoskeletal conditions. *J Occup Rehabil* 2020;30(3):318-330. [doi: [10.1007/s10926-019-09843-4](https://doi.org/10.1007/s10926-019-09843-4)] [Medline: [31267266](https://pubmed.ncbi.nlm.nih.gov/31267266/)]
75. Janssen M, Brous P, Estevez E, Barbosa LS, Janowski T. Data governance: organizing data for trustworthy artificial intelligence. *Gov Inf Q* 2020;37(3):101493. [doi: [10.1016/j.giq.2020.101493](https://doi.org/10.1016/j.giq.2020.101493)]
76. Liang W, Tadesse GA, Ho D, Fei-Fei L, Zaharia M, Zhang C, et al. Advances, challenges and opportunities in creating data for trustworthy AI. *Nat Mach Intell* 2022;4(8):669-677. [doi: [10.1038/s42256-022-00516-1](https://doi.org/10.1038/s42256-022-00516-1)]
77. Braithwaite J, Herkes J, Ludlow K, Testa L, Lamprell G. Association between organisational and workplace cultures, and patient outcomes: systematic review. *BMJ Open* 2017;7(11):e017708 [FREE Full text] [doi: [10.1136/bmjopen-2017-017708](https://doi.org/10.1136/bmjopen-2017-017708)] [Medline: [29122796](https://pubmed.ncbi.nlm.nih.gov/29122796/)]
78. Shanafelt TD, Gorringer G, Menaker R, Storz KA, Reeves D, Buskirk SJ, et al. Impact of organizational leadership on physician burnout and satisfaction. *Mayo Clin Proc* 2015;90(4):432-440. [doi: [10.1016/j.mayocp.2015.01.012](https://doi.org/10.1016/j.mayocp.2015.01.012)] [Medline: [25796117](https://pubmed.ncbi.nlm.nih.gov/25796117/)]
79. Xueyun Z, Al Mamun A, Masukujjaman M, Rahman MK, Gao J, Yang Q. Modelling the significance of organizational conditions on quiet quitting intention among Gen Z workforce in an emerging economy. *Sci Rep* 2023;13(1):15438 [FREE Full text] [doi: [10.1038/s41598-023-42591-3](https://doi.org/10.1038/s41598-023-42591-3)] [Medline: [37723179](https://pubmed.ncbi.nlm.nih.gov/37723179/)]
80. Ali Shah SA, Uddin I, Aziz F, Ahmad S, Al-Khasawneh MA, Sharaf M. An enhanced deep neural network for predicting workplace absenteeism. *Complexity* 2020;2020:1-12. [doi: [10.1155/2020/5843932](https://doi.org/10.1155/2020/5843932)]
81. Su TH, Wu CH, Kao JH. Artificial intelligence in precision medicine in hepatology. *J Gastroenterol Hepatol* 2021;36(3):569-580. [doi: [10.1111/jgh.15415](https://doi.org/10.1111/jgh.15415)] [Medline: [33709606](https://pubmed.ncbi.nlm.nih.gov/33709606/)]
82. Schafer KM, Kennedy G, Gallyer A, Resnik P. A direct comparison of theory-driven and machine learning prediction of suicide: a meta-analysis. *PLoS One* 2021;16(4):e0249833 [FREE Full text] [doi: [10.1371/journal.pone.0249833](https://doi.org/10.1371/journal.pone.0249833)] [Medline: [33844698](https://pubmed.ncbi.nlm.nih.gov/33844698/)]
83. Sallam M. ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. *Healthcare (Basel)* 2023;11(6):887 [FREE Full text] [doi: [10.3390/healthcare11060887](https://doi.org/10.3390/healthcare11060887)] [Medline: [36981544](https://pubmed.ncbi.nlm.nih.gov/36981544/)]
84. Purgato M, Singh R, Acarturk C, Cuijpers P. Moving beyond a 'one-size-fits-all' rationale in global mental health: prospects of a precision psychology paradigm. *Epidemiol Psychiatr Sci* 2021;30:e63 [FREE Full text] [doi: [10.1017/S2045796021000500](https://doi.org/10.1017/S2045796021000500)] [Medline: [34632978](https://pubmed.ncbi.nlm.nih.gov/34632978/)]
85. Becker S, Miron-Shatz T, Schumacher N, Krocza J, Diamantidis C, Albrecht UV. mHealth 2.0: experiences, possibilities, and perspectives. *JMIR mHealth uHealth* 2014;2(2):e24 [FREE Full text] [doi: [10.2196/mhealth.3328](https://doi.org/10.2196/mhealth.3328)] [Medline: [25099752](https://pubmed.ncbi.nlm.nih.gov/25099752/)]
86. Elston DM. The novelty effect. *J Am Acad Dermatol* 2021;85(3):565-566. [doi: [10.1016/j.jaad.2021.06.846](https://doi.org/10.1016/j.jaad.2021.06.846)] [Medline: [34153390](https://pubmed.ncbi.nlm.nih.gov/34153390/)]
87. Gerke S, Minssen T, Cohen G. Ethical and legal challenges of artificial intelligence-driven healthcare. *Artif Intell Healthcare* 2020;295-336. [doi: [10.1016/b978-0-12-818438-7.00012-5](https://doi.org/10.1016/b978-0-12-818438-7.00012-5)]
88. Rodrigues R. Legal and human rights issues of AI: gaps, challenges and vulnerabilities. *J Responsible Technol* 2020;4:100005. [doi: [10.1016/j.jrt.2020.100005](https://doi.org/10.1016/j.jrt.2020.100005)]

89. Andraško J, Mesarčík M, Hamulák O. The regulatory intersections between artificial intelligence, data protection and cyber security: challenges and opportunities for the EU legal framework. *AI Soc* 2021;36(2):623-636. [doi: [10.1007/s00146-020-01125-5](https://doi.org/10.1007/s00146-020-01125-5)]
90. Schönberger D. Artificial intelligence in healthcare: a critical analysis of the legal and ethical implications. *Int J Law Inf Technol* 2019;27(2):171-203. [doi: [10.1093/braincomms/fcae242](https://doi.org/10.1093/braincomms/fcae242)] [Medline: [39051028](https://pubmed.ncbi.nlm.nih.gov/39051028/)]
91. Fischer L, Ehrlinger L, Geist V, Ramler R, Sobiezyk F, Zellinger W, et al. AI system engineering—key challenges and lessons learned. *MAKE* 2021;3(1):56-83. [doi: [10.3390/make3010004](https://doi.org/10.3390/make3010004)]
92. Ibrahim H, Liu X, Denniston AK. Reporting guidelines for artificial intelligence in healthcare research. *Clin Exp Ophthalmol* 2021;49(5):470-476. [doi: [10.1111/ceo.13943](https://doi.org/10.1111/ceo.13943)] [Medline: [33956386](https://pubmed.ncbi.nlm.nih.gov/33956386/)]

Abbreviations

AI: artificial intelligence

DL: deep learning

ML: machine learning

NLP: natural language processing

OSF: Open Science Framework

OSH: occupational safety and health

PICO: patient or population, intervention, comparison, and outcomes

PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews

RQ: research question

RTW: return-to-work

WHPP: workplace health promotion and prevention

Edited by JL Raisaro; submitted 09.10.23; peer-reviewed by M Ijaz, C Ordun; comments to author 12.12.23; revised version received 02.01.24; accepted 10.07.24; published 20.08.24.

Please cite as:

Lange M, Löwe A, Kayser I, Schaller A

Approaches for the Use of AI in Workplace Health Promotion and Prevention: Systematic Scoping Review

JMIR AI 2024;3:e53506

URL: <https://ai.jmir.org/2024/1/e53506>

doi: [10.2196/53506](https://doi.org/10.2196/53506)

PMID: [38989904](https://pubmed.ncbi.nlm.nih.gov/38989904/)

©Martin Lange, Alexandra Löwe, Ina Kayser, Andrea Schaller. Originally published in JMIR AI (<https://ai.jmir.org>), 20.08.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR AI, is properly cited. The complete bibliographic information, a link to the original publication on <https://www.ai.jmir.org/>, as well as this copyright and license information must be included.

Review

Exploring Machine Learning Applications in Pediatric Asthma Management: Scoping Review

Tanvi Ojha^{1,2}, BSc; Atushi Patel¹, HBSc; Krishihan Sivapragasam¹, MSc; Radha Sharma^{1,2}, HBSc; Tina Vosoughi¹, HBSc; Becky Skidmore³, MLS; Andrew D Pinto^{1,4,5,6}, MD, MSc; Banafshe Hosseini^{1,5,6}, MSc, PhD

¹Upstream Lab, MAP Centre for Urban Health Solutions, Li Ka Shing Knowledge Institute, St. Michael's Hospital, Toronto, ON, Canada

²Temerty Faculty of Medicine, University of Toronto, Toronto, ON, Canada

³Independent Information Specialist, Ottawa, ON, Canada

⁴Department of Family and Community Medicine, St. Michael's Hospital, Toronto, ON, Canada

⁵Department of Family and Community Medicine, Temerty Faculty of Medicine, University of Toronto, Toronto, ON, Canada

⁶Division of Clinical Public Health & Institute for Health Policy, Management and Evaluation, Dalla Lana School of Public Health, University of Toronto, Toronto, ON, Canada

Corresponding Author:

Banafshe Hosseini, MSc, PhD

Upstream Lab, MAP Centre for Urban Health Solutions

Li Ka Shing Knowledge Institute

St. Michael's Hospital

30 Bond Street

Toronto, ON, M5B 1W8

Canada

Phone: 1 416 864 6060 ext 76148

Email: benita.hosseini@unityhealth.to

Abstract

Background: The integration of machine learning (ML) in predicting asthma-related outcomes in children presents a novel approach in pediatric health care.

Objective: This scoping review aims to analyze studies published since 2019, focusing on ML algorithms, their applications, and predictive performances.

Methods: We searched Ovid MEDLINE ALL and Embase on Ovid, the Cochrane Library (Wiley), CINAHL (EBSCO), and Web of Science (core collection). The search covered the period from January 1, 2019, to July 18, 2023. Studies applying ML models in predicting asthma-related outcomes in children aged <18 years were included. Covidence was used for citation management, and the risk of bias was assessed using the Prediction Model Risk of Bias Assessment Tool.

Results: From 1231 initial articles, 15 met our inclusion criteria. The sample size ranged from 74 to 87,413 patients. Most studies used multiple ML techniques, with logistic regression (n=7, 47%) and random forests (n=6, 40%) being the most common. Key outcomes included predicting asthma exacerbations, classifying asthma phenotypes, predicting asthma diagnoses, and identifying potential risk factors. For predicting exacerbations, recurrent neural networks and XGBoost showed high performance, with XGBoost achieving an area under the receiver operating characteristic curve (AUROC) of 0.76. In classifying asthma phenotypes, support vector machines were highly effective, achieving an AUROC of 0.79. For diagnosis prediction, artificial neural networks outperformed logistic regression, with an AUROC of 0.63. To identify risk factors focused on symptom severity and lung function, random forests achieved an AUROC of 0.88. Sound-based studies distinguished wheezing from nonwheezing and asthmatic from normal coughs. The risk of bias assessment revealed that most studies (n=8, 53%) exhibited low to moderate risk, ensuring a reasonable level of confidence in the findings. Common limitations across studies included data quality issues, sample size constraints, and interpretability concerns.

Conclusions: This review highlights the diverse application of ML in predicting pediatric asthma outcomes, with each model offering unique strengths and challenges. Future research should address data quality, increase sample sizes, and enhance model interpretability to optimize ML utility in clinical settings for pediatric asthma management.

(JMIR AI 2024;3:e57983) doi:[10.2196/57983](https://doi.org/10.2196/57983)

KEYWORDS

pediatric asthma; machine learning; predictive modeling; asthma management; exacerbation; artificial intelligence

Introduction

Background

Asthma is characterized by inflammation and narrowing of the airways, leading to recurring episodes of wheezing, breathlessness, coughing, and chest tightness. As the most prevalent chronic childhood condition, asthma affects approximately 14% of children worldwide [1,2] and ranks among the top conditions for disability-adjusted life years in children [3]. Severe asthma exacerbations, defined as those requiring systemic corticosteroids, emergency department (ED) visits, or hospitalization, are not only the primary cause of urgent health care visits, hospitalizations, and asthma-related mortality in children but contribute to asthma-related morbidity and mortality in children, incurring substantial treatment costs [4,5].

Risk factors for asthma exacerbations are multifaceted, ranging from socioeconomic factors to environmental exposures. Low income, residing in areas of concentrated poverty, limited access to health care providers, and high medication costs are significant contributors [6-8]. In addition, factors such as systemic and interpersonal racial and ethnic discrimination, suboptimal asthma control, and environmental triggers play a crucial role in exacerbation development [9,10]. Specifically, aeroallergen exposure or sensitization and concurrent viral infections have been shown to significantly increase exacerbation risks [11-13]. Given this complex interplay of factors, accurately predicting severe asthma exacerbations in children remains a challenge. Accurate prediction of children at risk for severe exacerbations can facilitate preemptive care strategies, reduce morbidity, and enhance the quality of life of those affected [14].

Machine learning (ML), a branch of artificial intelligence (AI), emerges as a promising tool. A range of supervised learning techniques, such as linear and logistic regression, decision trees, and classifier methods, including support vector machines (SVMs) and gradient boosting, are used to predict specific data categories (eg, asthmatic vs nonasthmatic) or continuous variables (eg, lung function measurements) [15]. In contrast, unsupervised learning techniques, such as k-means clustering and hierarchical clustering, are used to develop models that enable the clustering of the data [15]. ML's ability to analyze data and identify patterns has already shown success in various medical applications, including electrocardiography interpretation, heart failure classification, and diabetes outcome prediction [16-18]. In asthma management, AI has been instrumental in diagnosis, severity classification, and even in predicting asthma-related hospitalization risks at emergency encounters [19-22]. Several studies have investigated the role of AI in monitoring asthma exacerbations. Real-time assessment tools using environmental and physiological sensors have demonstrated notable accuracy in predicting exacerbations [23]. Contactless bed sensors for nocturnal data collection have also shown promise in detecting exacerbations [24]. In addition, AI-assisted clinical decision support tools, such as the Asthma

Guidance and Prediction System, have been evaluated for their efficacy in reducing exacerbation frequency in children [25].

Recent advancements in ML offer promising tools for predicting asthma exacerbations. A previous systematic review highlighted the moderate predictive performance of traditional models, with emerging ML approaches showing potential for enhancing prediction accuracy [26]. Similarly, another recent systematic review and meta-analysis of 11 studies, focusing on participants aged ≥ 5 years with preexisting asthma diagnoses, demonstrated good discrimination. The overall pooled area under the receiver operating characteristic curve (AUROC) was 0.80 (95% CI 0.76-0.83), and the diagnostic odds ratio was 7.02 (95% CI 5.20-9.47), indicating that ML-based prediction models for asthma exacerbation could achieve substantial accuracy [27]. Notably, of the 11 studies included in the 2022 systematic review, 6 (55%) were conducted after 2019, indicating considerable advancements in a short period [27]. However, these studies focused on participants aged > 5 years, leaving a gap in research for younger children [27]. Therefore, our scoping review aims to focus exclusively on studies conducted since 2019 that applied ML in predicting asthma exacerbations in children aged < 18 years.

Objectives

We intend to consolidate current knowledge by examining recent studies. This includes describing the types of predictive models developed, their applications in various settings, and the populations targeted and evaluating their performance in terms of accuracy, sensitivity, and specificity. This targeted approach will provide insights into the latest ML advancements and their potential to enhance pediatric asthma care.

Methods

Search Strategy

We registered this systematic review with PROSPERO (CRD42023440928) and have used the PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) to guide our reporting.

Search Strategy and Eligibility Criteria

An experienced information specialist (BS) developed and tested the search strategies in an iterative process in consultation with the review team. The MEDLINE strategy was peer reviewed by another senior information specialist before execution using the Peer Review of Electronic Search Strategies checklist [28]. Using the multfile and deduplication tool available on the Ovid platform, we searched Ovid MEDLINE ALL and Embase Classic+Embase. We also searched the Cochrane Library (Wiley), CINAHL (EBSCO), and Web of Science (core collection). All searches were performed on July 18, 2023. In addition, the reference lists of retrieved articles and relevant reviews were searched to identify other relevant studies.

The strategies used a combination of controlled vocabulary (eg, "Asthma," "Artificial Intelligence," and "Risk Assessment")

and keywords (eg, asthma, deep learning, and prognosis). There were no language restrictions on any of the searches, but results were limited to the publication years 2019 to the present. When possible, animal-only records, opinion pieces, and other irrelevant publication types (eg, case studies and conferences) were removed (refer to [Multimedia Appendix 1](#) for strategies). Records were downloaded and deduplicated using EndNote (version 9.3.3; Clarivate Analytics) and uploaded to Covidence (Veritas Health Innovation [29]) for efficient data management, extraction, and synthesis.

All studies were required to meet the eligibility criteria concerning the research focus, at both title/abstract and full-text screening: (1) in-vivo studies (human-based) that applied ML techniques to predict asthma-related outcomes, (2) participants aged <18 years, and (3) reported original data. The inclusion criteria were not limited to any specific study design to ensure inclusivity; hence, all available evidence from any study design was captured. There were no language restrictions for the studies reviewed. Studies were excluded if they were (1) in vitro studies (conducted on cellular substrates); (2) not focused on ML techniques to predict asthma-related outcomes; and (3) reviews, systematic reviews, opinions, editorials, and/or case reports.

Data Collection

Covidence was used throughout the review to manage citations. We engaged and trained several individuals to assist with reviewing citations (AP, RS, TO, and TV). During both parts of the screening process, the reviewers used the eligibility criteria to evaluate and determine the inclusion or exclusion of studies, which were then reported in Covidence. The first-level screening consisted of title and abstract screening of all uploaded studies. Each citation was reviewed by 2 people independently to select studies for full-text review (RS and TO). If the eligibility criteria were met completely, as assessed by both reviewers, the studies were included. If studies did not meet eligibility criteria, as determined by both reviewers, they were excluded. Any citations in which there was a difference in opinion were brought to the study team to discuss, and a third reviewer decided on inclusion or exclusion (AP and TV). Second-level screening involved a thorough assessment of all the studies that passed the initial screening on the basis of their title and abstracts, performed independently by 2 reviewers (RS and TO). An additional second-level review was performed by a solo reviewer (AP), who excluded any studies that did not meet the same eligibility criteria in the primary step and were considered ineligible. The final set of studies included in this scoping review includes only those that passed the full-text screening process. Two members of the study team (RS and TO) independently assisted with data extraction, with each study being extracted once. Subsequently, a comparison check was performed on each extracted study by a third reviewer (AP).

The following data were extracted: authors, title, journal, publication year, funding source, ML application types, the intended purpose of ML application, identification of any potential bias in the ML model design (if applicable), bias mitigation strategies (if applicable), study design, research question/study objective, primary and secondary outcomes, country, demographics, sample size, youth age groups, the unit

of analysis (individuals, groups, etc), data source (electronic medical records, databases, claims data, and health surveys), results, limitations, future research requirements (if applicable), use for clinical applications, and performance metrics (regression and classification). We noted if the information from an article was unavailable. A summary of the extracted information was recorded in Table S1 in [Multimedia Appendix 2](#) [25,30-43].

Risk of Bias Assessment

To assess the risk of bias, we used the Prediction Model Risk of Bias Assessment Tool (PROBAST) [44] and the guidelines for developing and reporting ML predictive models in biomedical research [45].

Data Synthesis

In this review, we used a narrative synthesis to thoroughly review and summarize the objectives, ML algorithms, and clinical relevance of each study. We focused on how these studies used ML to predict asthma-related outcomes in children, detailing the different ML algorithms, such as random forests (RFs), logistic regression, and neural networks, that were used and how they were applied. We organized the studies using the ML techniques they used and gathered key performance measures, such as accuracy, sensitivity, and specificity for each one. We also noted studies that used >1 ML method and identified and documented common limitations found within the studies, such as small sample sizes and generalizability issues.

Results

Study Selection and Characteristics

Our initial screening involved 1231 articles, from which 12 duplicates were removed using EndNote. This was followed by a primary screening that resulted in the inclusion of 102 studies. Upon secondary screening, 87 of these were excluded, leaving 15 articles that met our criteria for this review. The selection process is detailed in [Figure 1](#).

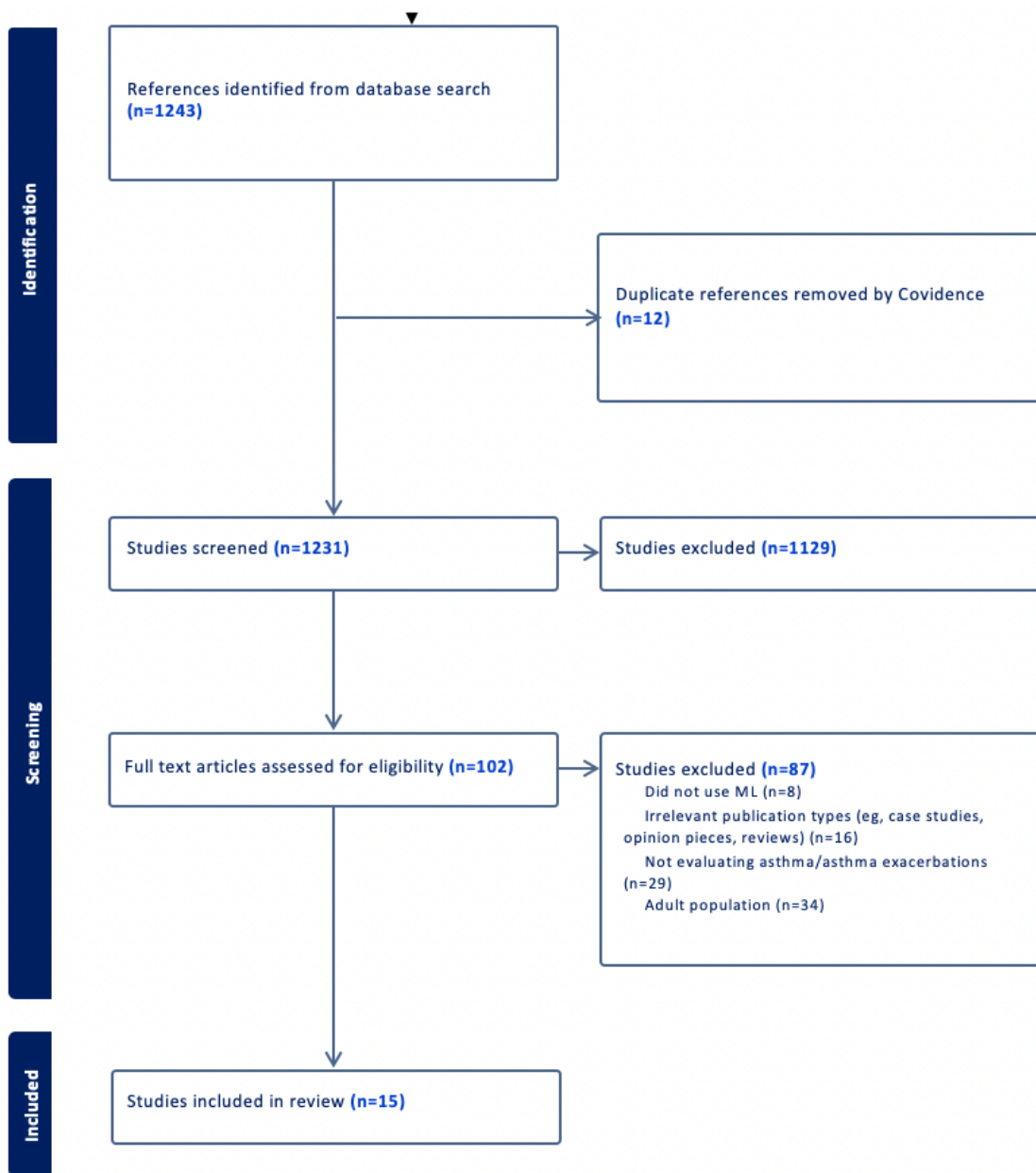
The included studies, published between 2019 and 2023, predominantly came out in 2021 [25,30-43]. They originated from various countries, including the United States (n=10, 67%) [25,30,32,34,35,38,39,41-43], Germany (n=1, 7%) [40], New Zealand (n=1, 7%) [31], Japan (n=1, 7%) [36], the United Kingdom (n=1, 7%) [33], and Singapore (n=1, 7%) [37]. Sample sizes in these studies ranged from 74 to 87,413 pediatric patients, indicating a wide variation in the population sizes examined.

Table S1 in [Multimedia Appendix 2](#) provides a comprehensive summary of the key data extracted from each included study. Most of these studies (n=9, 60%) implemented multiple ML techniques [30-34,38-40,43]. Logistic regression (n=7, 47%) and RFs (n=6, 40%) were the most commonly studied techniques [30-35,38-40,43]. This was followed by gradient boosting (n=4, 27%) [31,32,39,40] and artificial neural networks (ANNs; n=3, 20%) [30,38,41]. Decision trees (n=2, 13%) [34,36], natural language processing (NLP) models (n=2, 13%) [25,42], and Gaussian mixture models (n=1, 7%) [37] were the least frequent techniques used. Regarding study design,

retrospective cohort studies were predominant (n=9, 60%) [30-32,35,38,39,41-43], with a smaller proportion being prospective cohorts (n=5, 33%) [33,34,36,37,40] and a single randomized controlled trial (n=1, 7%) [25]. Detailed information

on the various ML models applied in the prediction of asthma exacerbations and related outcomes in children is provided in Tables S2-S8 in [Multimedia Appendix 2](#).

Figure 1. The selection process of eligible studies from all identified citations. ML: machine learning.



Quality Assessments

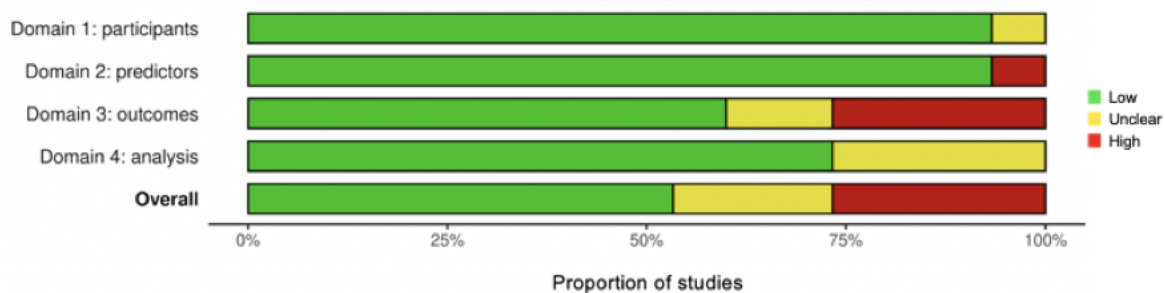
The risk of bias in the included studies was assessed using the PROBAST tool [44]. Our analysis revealed that most studies (n=8, 53%) exhibited a low risk of bias [30-32,34-36,40,41], indicating robust methodologies and reporting. However, some studies (n=3, 20%) were classified with an unclear risk [33,37,42] because of insufficient detail in certain aspects, whereas a few studies (n=4, 27%) were identified as high risk

[38,39,42,43], suggesting potential issues affecting their reliability. Studies classified as unclear or high risk often faced issues such as inconsistent definitions of outcomes across participants, outcome assessments influenced by prior knowledge of the predictors, or poorly specified inclusion and exclusion criteria for participants. Detailed breakdowns of each study’s bias assessment are presented in [Figure 2](#), and a summary of the overall risk across all studies is depicted in [Figure 3](#).

Figure 2. Risk of bias summary based on the Prediction Model Risk of Bias Assessment Tool quality assessment tool for included studies [25,30-43].

	Domain 1: participants	Domain 2: predictors	Domain 3: outcomes	Domain 4: analysis	Overall
Krautenbacher et al (2019)	Low	Low	Low	Low	Low
Bose et al (2021)	Low	Low	Low	Low	Low
Deng et al (2021)	Low	Low	Low	Low	Low
Sills et al (2021)	Low	Low	High	Low	High
Hurst et al (2021)	Low	Low	High	Low	High
Bhardwaj et al (2023)	Low	Low	Low	Low	Low
AlSaad et al (2022)	Low	Low	Low	Low	Low
Hogan et al (2022)	Low	Low	High	Unclear	High
Gorham et al (2023)	Low	Low	Low	Low	Low
Messinger et al (2019)	Low	Low	Low	Low	Low
Habukawa et al (2020)	Low	Low	Low	Low	Low
Seol et al (2021)	Low	Low	Unclear	Unclear	Unclear
Seol et al (2020)	Low	High	High	Low	High
Hee et al (2019)	Unclear	Low	Unclear	Unclear	Unclear
Deliu et al (2020)	Low	Low	Low	Unclear	Unclear

Figure 3. Summary of the risk of bias assessment.



ML Models in Pediatric Asthma: Predictive and Diagnostic Applications

Table 1 outlines the primary outcomes and the ML models used across the included studies. For predicting asthma exacerbations, the outcomes included any asthma-related health care encounter (outpatient visits, ED visits, and hospitalizations) or a prescription for a systemic steroid [25,30,35,38,39,43]. In classifying asthma phenotypes, the outcomes were the identification of allergic versus nonallergic asthma and the differentiation between mild and moderate-severe asthma [31,40,42]. For asthma diagnosis prediction, the outcomes were

the prediction of an asthma diagnosis and the calculation of a pediatric asthma score (PAS) [32,41]. Studies identifying potential risk factors for asthma-related outcomes focused on outcomes, including the severity of symptoms and lung function, considering factors such as family history, medical history, and environmental triggers [33,34]. In sound-based diagnosis studies, the outcomes included the identification of wheezing versus nonwheezing sounds and the differentiation between asthmatic and normal coughs [36,37]. Features commonly used across studies include demographic data, such as sex, age, and race, despite significant variations in ML models and outcomes [25,30,35,38,39,43].

Table 1. Application of ML^a models in pediatric asthma management through predictive and diagnostic modalities.

Category	Outcome	Primary ML models
Prediction of asthma exacerbations [25,30,35,38,39,43]	Any encounter (outpatients, ED ^b visits, and hospitalization) with an asthma-related ICD-9 or ICD-10 ^c code or a prescription for a systemic steroid	Neural networks, LASSO ^d regression, RFs ^e , XGBoost, and natural language processing
Classification of asthma phenotypes [31,40,42]	Allergic vs nonallergic asthma and mild vs moderate-severe asthma	SVMs ^f and stochastic gradient boosting
Asthma diagnosis prediction [32,41]	Prediction of asthma diagnosis and PAS ^g	XGBoost, ANNs ^h , and natural language processing
Identification of potential risk factors for asthma [33,34]	Potential risk factors (such as family hx ⁱ , medical hx, and environmental triggers) for asthma-related outcomes (including symptom severity and lung function)	K-means clustering, RFs, and decision tree
Sound-based asthma or wheezing diagnosis [36,37]	Identification of wheezing vs nonwheezing sounds and differentiation between asthmatic and normal coughs	Decision trees and Gaussian mixture models

^aML: machine learning.

^bED: emergency department.

^cICD-9 or ICD-10: *International Classification of Diseases*, 9th or 10th revisions.

^dLASSO: least absolute shrinkage and selection operator.

^eRF: random forest.

^fSVM: support vector machine.

^gPAS: pediatric asthma score.

^hANN: artificial neural network.

ⁱhx: history.

Table 2 provides a detailed summary of the predictors, clinical outcomes, and models used in the included studies. Studies have consistently used demographic data to predict asthma exacerbations. However, features related to medical history and health care use varied across the studies. Some studies focused on prescribed inhaled or oral steroids, previous health care use, and presence of moderate to severe asthma [25,30,35,39]. In contrast, others included variables such as time to triage, time to first medication and asthma medication, ED hourly volume, and patient disposition, including admitted or discharged [43]. Notably, some studies incorporated hospital characteristics, such as ownership (private vs public sector), teaching status, and size, along with family history factors such as alcohol or drug issues or housing instability [38]. Health insurance presence and type were also examined [39]. The models used in these studies included neural networks, least absolute shrinkage and selection operator regression, RFs, XGBoost, and NLP. The models were evaluated using metrics such as AUROC, accuracy, F₁-score, precision, recall, and specific measures such as mean average negative predictive value (NPV). The best-performing models varied by application. Recurrent neural networks [30] and XGBoost showed high performance in predicting asthma exacerbations, with XGBoost achieving an AUROC of 0.761 [39]. ANNs outperformed logistic regression in predicting hospital readmissions, achieving an AUROC of 0.637 [38]. RFs

were particularly effective in predicting hospitalization needs, with an AUROC of 0.886 [43].

A variety of demographics and clinical characteristics were used to differentiate between allergic and nonallergic asthma [31,40,42]. Key demographic variables included age, sex, weight, and race. Clinical parameters such as C-reactive protein levels, eosinophilic granulocytes, and oxygen saturation were also included in some studies [31]. Genetic markers, specifically protein kinase N2 and protein tyrosine kinase 2, along with breastfeeding duration, were also evaluated for their roles in asthma phenotypes [40]. In addition, some studies evaluated risk factors such as home conditions (eg, presence of carpets, home location and year, and animal triggers) and school characteristics, and home-related ventilators were considered to assess indoor environmental impacts on asthma [34]. ML models (eg, RFs, SVMs, gradient boosting, and decision trees) were used to analyze these variables. The most effective models varied across studies. Metrics such as AUROC, accuracy, precision, true positive rate, true negative rate, F₁-score, prevalence ratios, and IQRs were used to evaluate the models' performance. SVMs demonstrated high performance with metrics, including an accuracy of 77.8%, precision of 0.81, and an AUROC of 0.79. Stochastic gradient boosting achieved an AUROC of 0.81, highlighting its efficacy in incorporating genetic markers and breastfeeding duration.

Table 2. Summary of the included studies on ML^a applications in pediatric asthma: predictors, clinical outcomes, and models.

Study	Potential predictors, variables of interests-grouped	Metrics	Data source	Outcomes	ML models
AlSaad et al [30], 2022	Demographic data, medication use, health service use, clinical parameters and characteristics (comorbid illnesses), and insurance information	AUROC ^b (0.85), AUC ^c -PR ^d (0.74), and F_1 -score (0.61)	EHRs ^e	Frequency of ED ^f use (number of visits made by pediatric patients during a 1-year predication window)	Deep learning: recurrent neural networks
Bhardwaj et al [31], 2023	Demographic data (age and weight) and clinical parameters and characteristics (C-reactive protein, eosinophilic granulocytes, oxygen saturation, premedication inhaled corticosteroid+long-acting β -2 agonist, other premedication, Pulmicort or celestamine during hospitalization, and azithromycin during hospitalization)	SVM ^g differentiated between allergic and nonallergic asthma most well: accuracy (77.8%), precision (0.81), true positive rate (0.73), true negative rate (0.81), F_1 -score (0.81), and AUROC (0.79); because of the imbalance between both groups, a stratified 10-fold cross-validation was used	EHRs	Classify predominantly allergic asthma and nonallergic asthma among preschool children	RFs ^h , extreme gradient boosting, SVMs, adaptive boosting, extra tree classifier, and logistic regression
Bose et al [32], 2021	Demographic data (race, sex, ethnicity, and language spoken), geographic location (state of residency at the time of their first asthma diagnosis), insurance information (Medicaid enrollment), care site information (place of service such as EDs or office visits and provider specialties at first asthma diagnosis), medical hx ⁱ (age of first and last asthma diagnoses and nonasthma-related clinical visits)	Mean ANSA, median ANSA, precision, recall, F_1 -score, and accuracy; XGBoost presented the best mean ANSA ^j : mean ANSA (0.43), median ANSA (0.43), precision (0.95), recall (0.82), F_1 -score (0.88), and accuracy (0.81)	EHRs	Occurrence of asthma diagnosis by the age of 10 years following an asthma incident	Naive Bayes, K-nearest neighbors, logistic regression, RFs, and XG-Boost
Deliu et al [33], 2020	Medical hx and medication use (asthma diagnosis, use of asthma medication, current wheeze, asthma severity, and lung function) and risk factors (environmental tobacco smoke, pet ownership, length of breastfeeding, day-care attendance, presence of older siblings, and family hx of asthma)	FVC ^k , FEV1 ^l , IE ^m , FE ⁿ (early-onset frequent exacerbations), IE (93.7%), and FE (6.3%); shorter duration of breastfeeding was the strongest risk factor. FEV1/FVC of FE group: 85.1% at 8 years old	EHRs and health surveys	Examine risk factors that result in asthma-related outcomes in late childhood	K-means clustering
Deng et al [34], 2021	Demographic data (sex, race, age, and grade), family hx (job status, health status and hx, and education), insurance information, and risk factors (home conditions, such as carpet in house, tile flooring, or home location and year, animal triggers, home-related ventilators, and school characteristics)	Percentage and PR; top contributing factors: asthma, family rhinitis hx (relative importance: 10.40%), plant pollen trigger (relative importance: 5.48%), and bedroom carpet (relative importance: 3.58%). Allergy-related symptoms: plant pollen trigger (relative importance: 10.88%), higher paternal education (relative importance: 7.33%), and bedroom carpet (relative importance: 5.28%)	Health surveys	Evaluating factors in indoor environments (home vs school) contributing to asthma and allergy-related symptoms	RFs and decision tree
Gorham et al [35], 2023	Demographic data (age, sex, and race) and medical hx and medication use (inhaled or oral steroid prescribed, ED visits in a year, moderate to severe asthma, and asthma-related primary care visits in a year)	AUROC; internal validation: 0.769. 10-fold cross-validation AUROC: 0.737	EHRs	ED visit because of asthma exacerbations (also known as AER ^o); asthma exacerbations: asthma-related emergency	Logistic regression
Habukawa et al [36], 2020	Audio features (wheeze sounds: frequency, intensity, and duration) and demographic data (age)	Sensitivity, specificity, PPV ^p , and NPV ^q ; sensitivity (100%), specificity (95.7%), PPV (90.3%), and NPV (100%)	EHRs	Identification of wheeze sounds vs nonwheeze sounds	Decision tree

Study	Potential predictors, variables of interests-grouped	Metrics	Data source	Outcomes	ML models
Hee et al [37], 2019	Demographic data (age, sex, race, and weight), clinical parameters and characteristics (temperature, respiratory rate, heart rate, and shortness of breath), audio features (cough sounds: mel-frequency cepstral coefficients and constant-Q cepstral coefficients), and medical hx (asthma, allergic rhinitis, and recurrent wheeze)	Sensitivity (82.81%) and specificity (84.76%)	EHRs and health surveys	Classify and differentiate asthmatic coughs from normal voluntary coughs	Gaussian mixture model-universal background model
Hogan et al [38], 2022	Demographic data (sex and age), insurance, family hx (family member with alcohol or drug issues, hx of abuse, housing instability, and foster care), clinical parameters and characteristics (LOS ^f , admission season, and chronic conditions), and hospital characteristics (hospital ownership, teaching status, and hospital size)	AUC; logistic regression (0.592) and ANNs ^s (0.637)	Claims data and biomedical databases	Asthma hospital readmission 180 days after hospital discharge	Logistic regression and ANNs
Hurst et al [39], 2022	Demographic data (age and sex), medical hx and medication use (comorbidities and prescribed asthma control plan), insurance, and health care use (inpatient admissions, ambulatory visits, and ED)	AUC at day 30, 90, and 180; LASSO ^l (0.753, 0.740, and 0.732), RFs (0.757, 0.747, and 0.729), and XGBoost (0.761, 0.752, and 0.739)	EHRs and biomedical databases	Predict the occurrence of asthma exacerbation; asthma exacerbation: any encounter with an asthma-related ICD-9 or -10 ^u code and a prescription for a systemic steroid	LASSO, RFs, and XGBoost
Krautenbacher et al [40], 2019	Clinical parameters and characteristics (genes, including <i>PKN2</i> ^v , <i>PTK2</i> ^w , and <i>ALPP</i> ^x , and breastfeeding), and demographic data (age and sex)	AUC; boosting was the best model for all data sets: 0.81	Health surveys and biomedical databases	Distinguish between healthy children, those with mild to moderate allergic asthma, and those with nonallergic asthma	LASSO, elastic net, RFs, and stochastic gradient boosting
Messinger et al [41], 2019	Demographic data (age, sex, and race) and medication use, medical hx, and medications (LOS, PAS ^y including vital sign data such as heart rate, respiratory rate, oxygen saturation, respiratory support, and medications)	Median absolute error; balanced set MAE ^z : 1.21	EHRs and biomedical databases	Use of vital sign data to predict the presence of asthma and to generate a novel pediatric-automated asthma score	ANNs
Seol et al [42], 2020	Demographic data (age, sex, ethnicity, and weight), family hx (asthma and smoking during pregnancy), medical hx (diagnosis of asthma, eczema, allergic rhinitis, eosinophilia, total IgE ^{aa} , asthma and associated outcomes such as persistent asthma, pertussis, pneumonia), and health care use (visits per year)	Percentage; NLP ^{ab} -PAC ^{ac} + / NLP-API ^{ad} +; 1614 (20%), NLP-PAC+ only: 954 (12%), NLP-API+ only: 105 (1%), and NLP-PAC- / NLP-API-: 5523 (67%); NLP-PAC and NLP-API; asthmatic children classified as NLP-PAC+ / NLP-API+ showed earlier onset asthma, more Th2 ^{ae} -high profile, poorer lung function, higher asthma exacerbation, and higher risk of asthma-associated comorbidities compared with other groups	EHRs	Identifying characteristics that will identify childhood asthma and its subgroups using 2 algorithms	NLP
Seol et al [25], 2021	Medical hx and medications (IgE count, eosinophil count, smoking exposure, hx of allergic rhinitis, previous exacerbations, asthma diagnosis, and medication use) and demographic data (age, sex, and race)	IQR and <i>P</i> value; asthma exacerbation: intervention 12%, control 15%, <i>P</i> =.60; Time (min) taken by the clinician to take a clinical decision, median: intervention 3.5 min vs control 11.3 min	EHRs	Determine the presence of asthma exacerbation to reduce its frequency using clinical information; asthma exacerbation: ED visit, hospitalization, or outpatient visit requiring systemic corticosteroids for asthma	NLP

Study	Potential predictors, variables of interests-grouped	Metrics	Data source	Outcomes	ML models
Sills et al [43], 2021	Demographic data (age, race, and sex), insurance, medical hx, and medications (ED and treatment factors: time to triage, time to first medication and asthma medication, ED hourly volume, and disposition including admitted or discharged)	AUC, accuracy, and F_1 -; model 1: triage (RF-AUC 0.831, accuracy 0.777, and F_1 -score 0.635, and logistic regression-AUC 0.795, accuracy 0.731, and F_1 -score 0.564); model 2: 60 minutes after patients' arrival (RF-AUC 0.886, accuracy 0.795, and F_1 -score 0.689, and logistic regression-AUC 0.823, accuracy 0.753, and F_1 -score 0.618)	EHRs	Predict the need for hospitalization of pediatric patients with asthma	RFs and logistic regression

^aML: machine learning.

^bAUROC: area under the receiver operating characteristic curve.

^cAUC: area under cover.

^dPR: precision recall.

^eEHR: electronic health record.

^fED: emergency department.

^gSVM: support vector machine.

^hRF: random forest.

ⁱhx: history.

^jANSA: average negative predictive value specificity area.

^kFVC: forced vital capacity.

^lFEV1: forced expiratory volume in the first second.

^mIE: infrequent exacerbation.

ⁿFE: frequent exacerbation.

^oAER: asthma emergency risk.

^pPPV: positive predictive value.

^qNPV: negative predictive value.

^rLOS: length of stay.

^sANN: artificial neural network.

^tLASSO: least absolute shrinkage and selection operator.

^uICD-9 or -10: International Classification of Diseases, 9th or 10th Revisions.

^vPKN2: protein kinase N2.

^wPTK2: protein tyrosine kinase 2.

^xALPP: alkaline phosphatase, placental.

^yPAS: pediatric asthma score.

^zMAE: masked autoencoder.

^{aa}IgE: immunoglobulin E.

^{ab}NLP: natural language processing.

^{ac}PAC: predetermined asthma criteria.

^{ad}API: Asthma Predictive Index.

^{ae}Th2: T helper 2 cells.

Studies that attempted to predict asthma diagnosis included a range of features, ML models, and metrics [32,41]. One study used demographic data such as race, sex, ethnicity, and language spoken, alongside medical history factors such as age at first and last asthma diagnoses and the number of nonasthma-related clinical visits, as well as geographic information such as the state of residency at the time of the first asthma diagnosis and insurance details, including Medicaid enrollment [32]. Another study focused on using patients' medical history and medication use, along with vital sign data, to predict the presence of asthma

and generate a novel PAS [41]. Various ML models were used, including naive Bayes, k-nearest neighbors, logistic regression, RFs, ANNs, and XGBoost, with ANNs and XGBoost showing the best performance. The metrics used to evaluate these models included mean average NPV specificity area, median average NPV specificity area, precision, recall, F_1 -score, and accuracy.

To identify potential risk factors for asthma-related outcomes, particularly focusing on the severity of symptoms and lung function, various ML models were used [33,34]. One study

examined a range of variables, including medical history and medication use, such as asthma diagnosis, current wheeze, asthma severity, and lung function, alongside risk factors such as environmental tobacco smoke, pet ownership, length of breastfeeding, day-care attendance, presence of older siblings, and family history of asthma. K-means clustering was used to identify patterns and categorize risk factors associated with different asthma outcomes [33]. Evaluation metrics included forced vital capacity and forced expiratory volume in the first second, with specific attention to infrequent exacerbations and early-onset frequent exacerbations. Shorter breastfeeding duration emerged as the strongest risk factor, with the forced expiratory volume in the first second/forced vital capacity ratio in the frequent exacerbation group being 85.1% at 8 years old [33]. Another study focused on demographic data, such as sex, race, age, and grade, along with family history variables, including job status, health status, and education [34]. The study also considered insurance information and risk factors such as home conditions (eg, presence of carpets or tile flooring and home location and year), animal triggers, home-related ventilators, and school characteristics. Using RFs and decision trees, the study identified key contributors to asthma and allergy-related symptoms. The metrics used included prevalence ratios. Significant factors for asthma included a family history of rhinitis (relative importance of 10.40%), plant pollen trigger (relative importance of 5.48%), and bedroom carpet (relative importance of 3.58%). For allergy-related symptoms, important factors were plant pollen trigger (relative importance of 10.88%), higher paternal education (relative importance of 7.33%), and bedroom carpet (relative importance of 5.28%) [34].

To identify and classify asthmatic sounds, particularly focusing on wheezing and cough patterns, various ML models were used through a combination of audio features, demographic, and clinical data [36,37]. One study focused on differentiating between wheezing and nonwheezing sounds using a decision tree model [36]. The key features analyzed included audio characteristics such as the frequency, intensity, and duration of wheezing sounds, along with demographic data such as age. The model's performance was evaluated using metrics such as sensitivity, specificity, positive predictive value, and NPV. The decision tree model achieved a sensitivity of 100%, specificity of 95.7%, positive predictive value of 90.3%, and NPV of 100%, demonstrating its high accuracy in identifying wheezing sounds among pediatric patients [36]. Another study aimed to classify and differentiate asthmatic coughs from normal voluntary coughs using a Gaussian mixture model-universal background model [37]. This study incorporated audio features such as mel-frequency cepstral coefficients and constant-Q cepstral coefficients, along with demographic data (age, sex, race, and weight) and clinical parameters (temperature, respiratory rate, heart rate, and shortness of breath). In addition, medical history factors such as asthma, allergic rhinitis, and recurrent wheezing were included. The model's effectiveness was measured using sensitivity and specificity, achieving sensitivity of 82.81% and specificity of 84.76% [37]. These metrics indicate the model's robustness in accurately classifying asthmatic coughs and distinguishing them from normal coughs.

Common Limitations in the Reviewed Studies

A recurring theme in the limitations reported by the included studies pertains to challenges with data quality and completeness. Issues such as missing, incomplete, or limited data availability from medical records and health surveys were highlighted in several studies [34,38,41-43]. These data constraints can significantly impact the robustness and generalizability of the study findings. In the context of predicting asthma exacerbations, 3 studies specifically cited deficiencies in electronic health records (EHRs) [30,41,42] and pointed out the lack of critical variables in EHRs, such as socioeconomic status and adherence to treatment. These deficiencies arose from variables not being commonly recorded in EHRs. The absence of these variables can limit the depth and accuracy of predictive modeling, thereby affecting the models' performance and generalizability. Another notable limitation was the issue of imbalanced data sets [30-32], which refers to situations where the number of observations in different classes is disproportionately distributed. For example, if there are significantly more cases of nonasthmatic patients compared to patients with asthma, this imbalance can lead to biased or skewed models that do not perform well across all classes. Small sample sizes, which can affect the statistical power and validity of the findings, were also a concern in a few studies [25,31,33,40]. A small sample size generally refers to a data set that is not large enough to yield statistically significant results or reliable conclusions. This can vary depending on the study design and statistical methods used, but typically, small sample sizes limit the ability to generalize findings to a larger population. In addition, limitations were identified in studies focusing on wheezing and asthmatic cough recognition algorithms. For example, a study developed a wheeze detection device for use in home environments, raising questions about its clinical value because of the specific context of its intended application [36]. Similarly, another study [37] on an asthmatic cough recognition algorithm highlighted that its validity and accuracy depended on the correct labeling of coughs by attending physicians. These limitations underscore the need for improved data quality and data collection processes to enhance the reliability and applicability of ML models in pediatric asthma research.

Discussion

Principal Findings

This scoping review successfully identified 15 peer-reviewed studies published since 2019, focusing on ML models in predicting pediatric asthma outcomes. Model use was diverse: logistic regression (7 studies), RFs (6 studies), gradient boosting (4 studies), ANNs (3 studies), decision trees (2 studies), NLP (2 studies), and Gaussian mixture model (1 study), with area under the curve ranging from 0.62 to 0.88. Most studies (n=8, 53%) had a low to moderate risk of bias, and they were evaluated using PROBAST.

Comparative Analysis of ML Models

Among traditional ML models, logistic regression has demonstrated robustness, particularly in predicting hospitalization needs in pediatric asthma cases [30-33,35,38,43].

However, comparing logistic regression to RFs reveals that the latter offers superior performance in certain scenarios. For instance, RFs exhibited a higher area under the curve at the 1-hour postarrival time point in predicting hospitalization needs [43].

Gradient boosting models, particularly XGBoost, showed promise in certain scenarios. For example, in predicting early childhood asthma persistence, XGBoost matched the accuracy of logistic regression [32]. However, these models still lag slightly behind logistic regression and RFs in classifying asthma types, highlighting the potential differences in model efficacy across various applications.

The application of ANN provided promising results in predicting ED visits and asthma readmissions [30,38]. However, their performance, especially in complex clinical settings, warrants additional exploration and comparison with more conventional models. Decision trees, applied in more niche areas such as environmental risk assessment and wheeze sound recognition, demonstrated high accuracy and specificity [34,36]. NLP models, used within EHRs, helped early identification of pediatric asthma criteria [25,42], and Gaussian mixture models were applied to differentiate between patients with asthma and nonasthmatic patients through auditory recognition of types of coughs [37].

Application of Predictive Models Across Different Outcomes

Among the 15 studies, key outcomes include predicting asthma exacerbations requiring urgent care, classifying asthma phenotypes by identifying allergic versus nonallergic asthma and severity levels, predicting asthma diagnoses and calculating PAS, and identifying potential risk factors such as symptom severity and lung function. In addition, sound-based diagnosis studies focused on distinguishing wheezing and differentiating asthmatic from normal coughs. One study [39] developed predictive models for pediatric asthma exacerbations using sociodemographic data, comorbidities, medication prescriptions, prescribed asthma controller plans, and patient service use history. This algorithm functioned as a potent tool capable of identifying children at risk of asthma exacerbations. Consequently, it signaled when preventive measures would be valuable to implement. Several studies used ML models to predict hospitalization needs and readmission risks using demographic variables. The studies by Sills et al [43] and Hogan et al [38] used ML models using varying features, including demographic variables such as sex, age, and race to predict hospitalization needs and readmission risks. Sills et al [43] demonstrated the potential of 2 distinct ML models to predict hospitalization in pediatric asthma cases, highlighting the models' utility as supportive tools for clinical decision-making.

Similarly, Hogan et al [38] used an ANN algorithm to predict asthma readmissions within 180 days after discharge, finding that ANN outperformed traditional models in identifying readmission predictors. AlSaad et al [30] and Gorham et al [35] conducted studies focusing on predicting ED visits using data from EHRs/electronic medical records. Notably, the studies found that increased access to primary care with regular follow-ups resulted in fewer ED visits, suggesting that more

frequent visits allowed for better assessment and management of asthma. Their findings suggest that ML models can effectively identify children with asthma who are at higher risk of repeated ED visits. Given the challenges associated with frequent ED use in emergency care, these prediction models emerge as valuable tools in enhancing asthma management and assisting in clinical decision-making.

We also examined the role of ML in asthma diagnosis in a pediatric population. One study [37] developed an ML model to distinguish between asthmatic and normal coughs by creating a database of cough sounds from asthmatic and nonasthmatic children. Another study [36] focused on an ML-based wheeze detection algorithm, analyzing lung sounds recorded through stethoscopes. Both these studies exemplify the use of ML in identifying asthma symptoms accurately. In addition, an ML algorithm was explored to automate asthma severity scoring, aiming to create a pediatric asthma respiratory score from vital sign data [41]. Additional research [42] used an NLP model to identify asthma early in children, and another study [25] developed the Asthma Guidance and Prediction System using ML and NLP to enhance asthma management programs and reduce asthma exacerbations. These studies collectively demonstrate the considerable potential of ML in improving the diagnosis, severity assessment, and management of pediatric asthma.

In examining asthma phenotypes, several studies have leveraged ML to categorize different characteristics of asthma. Two studies implemented various ML techniques [31,32], focusing on EHR data to classify asthma types. One study [31] aimed to distinguish between allergic and nonallergic asthma, whereas another study [32] sought to predict persistent versus transient asthma. Similarly, 2 studies [25,42] used EHR data and applied an NLP algorithm to identify pediatric asthma subgroups. This capability to distinguish between different types of asthma can significantly inform clinical decisions and guide parents in choosing appropriate asthma treatments, as highlighted by others [32].

Further support for the use of ML in understanding asthma phenotypes and allergies comes from the studies of Deng et al [34] and Krautenbacher et al [40], each adopting a unique approach. Deng et al [34] used ML models to assess risk factors in home and school environments affecting asthma and allergies. In contrast, Krautenbacher et al [40] developed a unique ML method to enhance the prediction of childhood asthma phenotypes, specifically distinguishing between allergic and nonallergic asthma, using various inputs such as genotypes, questionnaires, and diagnostic tools. Both studies effectively demonstrated the potential of ML models in identifying asthma and allergy risk factors as well as in improving the classification of childhood asthma types. Similarly, another study [33] applied ML to analyze wheeze exacerbation trajectories in children using medical record data, revealing diverse exacerbation patterns, early life risk factors, and asthma outcomes. This study aligns with the others in using ML to discern patterns predictive of childhood asthma. Jeddi et al [46] further emphasize the significance of these findings, noting that the ability to identify factors associated with childhood asthma via ML can help predict children considered susceptible. This prediction, in turn,

enables the implementation of targeted interventions to prevent the onset of the disease.

Future Directions and Key Considerations

Applying ML models to predict asthma outcomes in children involves several critical considerations to ensure accuracy, reliability, and applicability. The basis of any ML model is the data it is trained on. It should be comprehensive and include variables such as age, sex, family medical history, environmental exposures (such as allergens, pollutants, and community viral loads), lifestyle factors (diet and physical activity), and clinical data (symptoms, medication use, lung function tests, etc). Several studies highlighted missing or incomplete data in medical records and health surveys [34,38,41-43], which underscores the importance of robust strategies for handling such data challenges. For example, studies have demonstrated that simple imputation methods, considering informative missingness, can be effective in managing missing numerical data in EHR for ML [47]. In addition, research on imputing missing values in laboratory data from EHRs has shown that the pattern of missingness is typically nonrandom and closely related to patients' comorbidities, suggesting that multilevel imputation algorithms are more effective than cross-sectional methods [48].

Another point to consider is that asthma is a chronic condition with variable progression over time. Incorporating longitudinal data, which means tracking patient data over time, can help the model recognize patterns and predict future exacerbations or improvements. In addition, there is limited information on the choice of ML models across different age groups within the pediatric population. This gap highlights the need for future research to specifically address the performance and applicability of ML models in different pediatric age groups. This approach could provide valuable insights into age-specific predictive features and model adjustments.

Beyond accuracy, the model must also be interpretable [49]. Clinicians and patients should be able to understand how and why a particular prediction was made, which builds trust and ensures that the model's findings are useful in real-world clinical decision-making. The model should also integrate seamlessly into existing clinical workflows. This involves considering how predictions will be delivered and their impact on clinical decision-making and ensuring they are in a format that health care providers can understand and easily incorporate into their existing decision-making processes. Previous research has shown that user-centered design is essential for successful implementation. For instance, a study involving 14 clinicians highlighted the need to identify patients at high risk and take proactive measures to manage asthma effectively [50]. Clinicians emphasized the importance of clear, actionable insights from the tool and understanding the underlying reasons for predictions. Barriers to implementation included usability

and workflow integration challenges; the need for clear algorithm explainability; and ensuring the tool's acceptability, adoption, and sustainability through proper design and training [50]. By involving clinicians in the design process, the tool was tailored to meet their needs, which underscores the importance of user-centered design in developing effective clinical decision support tools.

Strengths of this review included a comprehensive and systematic search across multiple databases, along with establishing clearly defined inclusion and exclusion criteria. The structured study selection process added robustness to the review. In addition, the use of the PROBAST tool for risk of bias assessment augmented the credibility of the review [44]. However, the review also had limitations that should be acknowledged. Despite a broad and inclusive search strategy designed to capture all subtypes of ML related to childhood asthma, some relevant studies might not be published in the indexed journals included in our search databases, and thus, there remains a possibility that some pertinent articles may have been inadvertently excluded.

This review highlights the potential of ML in transforming pediatric asthma care, from predicting exacerbations to characterizing asthma types. However, it also underscores the need for improved data quality, larger and more balanced data sets, and more rigorous validation to ensure these tools are clinically valuable. The exploration of varied ML techniques across studies offers a road map for future research to build more accurate, reliable, and applicable models for pediatric asthma management.

Conclusions

This scoping review provides a broad overview of ML applications used to predict asthma-related outcomes in children. We reviewed a diverse range of studies focused on the design, training, testing, and interpretation of ML models and observed that using ML in childhood asthma is an emerging field that has seen significant growth over the past few years. This recent surge in research highlights the evolving nature and increasing interest in applying ML to improve pediatric asthma outcomes.

By leveraging data from multiple sources, ML approaches have made strides in identifying distinct asthma phenotypes, paving the way for more tailored and effective treatment strategies in clinical practice. However, the field faces ongoing challenges, particularly regarding minimizing missing data, ensuring robust model validation, and achieving interpretability. In addition, integrating these models smoothly into clinical workflows remains a key obstacle. While ML holds considerable promise in pediatric asthma research, the field is still evolving. To fully realize its potential, further research is needed to address these challenges and enhance the practical application of ML models in clinical settings.

Acknowledgments

The authors thank Kaitryn Campbell (master of library and information science and master of science) for peer review of the MEDLINE search strategy and Roxana Rabet for assisting in the submission process.

Data Availability

The data sets generated and analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

BH and ADP conceived the study. TO, RS, AP, TV, and KS conducted the scoping review and drafted the first version of the manuscript. BS developed and executed the search strategy. BH supervised the conduct of this review from inception to data extraction and prepared the final version of the manuscript. All authors contributed to revising the manuscript for important intellectual content, provided final approval of the version to be published, and agreed to be accountable for all aspects of the work.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Search strategy.

[[DOCX File , 35 KB - ai_v3i1e57983_app1.docx](#)]

Multimedia Appendix 2

Study details.

[[DOCX File , 133 KB - ai_v3i1e57983_app2.docx](#)]

Multimedia Appendix 3

PRISMA-ScR checklist.

[[DOCX File , 106 KB - ai_v3i1e57983_app3.docx](#)]

References

1. Martin J, Townshend J, Brodlie M. Diagnosis and management of asthma in children. *BMJ Paediatr Open* 2022 Apr 26;6(1):e001277 [[FREE Full text](#)] [doi: [10.1136/bmjpo-2021-001277](https://doi.org/10.1136/bmjpo-2021-001277)] [Medline: [35648804](#)]
2. Zar HJ, Ferkol TW. The global burden of respiratory disease-impact on child health. *Pediatr Pulmonol* 2014 May 09;49(5):430-434. [doi: [10.1002/ppul.23030](https://doi.org/10.1002/ppul.23030)] [Medline: [24610581](#)]
3. Vos T, Flaxman AD, Naghavi M, Lozano R, Michaud C, Ezzati M, et al. Years lived with disability (YLDs) for 1160 sequelae of 289 diseases and injuries 1990-2010: a systematic analysis for the Global Burden of Disease Study 2010. *Lancet* 2012 Dec 15;380(9859):2163-2196 [[FREE Full text](#)] [doi: [10.1016/S0140-6736\(12\)61729-2](https://doi.org/10.1016/S0140-6736(12)61729-2)] [Medline: [23245607](#)]
4. Dharmage SC, Perret JL, Custovic A. Epidemiology of asthma in children and adults. *Front Pediatr* 2019 Jun 18;7:246 [[FREE Full text](#)] [doi: [10.3389/fped.2019.00246](https://doi.org/10.3389/fped.2019.00246)] [Medline: [31275909](#)]
5. Reddel HK, Taylor DR, Bateman ED, Boulet LP, Boushey HA, Busse WW, et al. An official American Thoracic Society/European Respiratory Society statement: asthma control and exacerbations: standardizing endpoints for clinical asthma trials and clinical practice. *Am J Respir Crit Care Med* 2009 Jul 01;180(1):59-99. [doi: [10.1164/rccm.200801-060st](https://doi.org/10.1164/rccm.200801-060st)]
6. Williams DR, Sternthal M, Wright RJ. Social determinants: taking the social context of asthma seriously. *Pediatrics* 2009 Mar;123 Suppl 3(Suppl 3):S174-S184 [[FREE Full text](#)] [doi: [10.1542/peds.2008-2233H](https://doi.org/10.1542/peds.2008-2233H)] [Medline: [19221161](#)]
7. Dubaybo BA. The care of asthma patients in communities with limited resources. *Res Rep Trop Med* 2021;12:33-38 [[FREE Full text](#)] [doi: [10.2147/RRTM.S247716](https://doi.org/10.2147/RRTM.S247716)] [Medline: [33727880](#)]
8. Karaca-Mandic P, Jena AB, Joyce GF, Goldman DP. Out-of-pocket medication costs and use of medications and health care services among children with asthma. *JAMA* 2012 Mar 28;307(12):1284-1291 [[FREE Full text](#)] [doi: [10.1001/jama.2012.340](https://doi.org/10.1001/jama.2012.340)] [Medline: [22453569](#)]
9. Thakur N, Barcelo NE, Borrell LN, Singh S, Eng C, Davis A, et al. Perceived discrimination associated with asthma and related outcomes in minority youth: the GALA II and SAGE II studies. *Chest* 2017 Apr;151(4):804-812 [[FREE Full text](#)] [doi: [10.1016/j.chest.2016.11.027](https://doi.org/10.1016/j.chest.2016.11.027)] [Medline: [27916618](#)]
10. Navanandan N, Hatoun J, Celedón JC, Liu AH. Predicting severe asthma exacerbations in children: blueprint for today and tomorrow. *J Allergy Clin Immunol Pract* 2021 Jul;9(7):2619-2626. [doi: [10.1016/j.jaip.2021.03.039](https://doi.org/10.1016/j.jaip.2021.03.039)] [Medline: [33831622](#)]
11. Kloefer KM, Gern JE. Virus/allergen interactions and exacerbations of asthma. *Immunol Allergy Clin North Am* 2010 Nov;30(4):553-63, vii [[FREE Full text](#)] [doi: [10.1016/j.iac.2010.08.002](https://doi.org/10.1016/j.iac.2010.08.002)] [Medline: [21029938](#)]
12. Friedlander SL, Busse WW. The role of rhinovirus in asthma exacerbations. *J Allergy Clin Immunol* 2005 Aug;116(2):267-273. [doi: [10.1016/j.jaci.2005.06.003](https://doi.org/10.1016/j.jaci.2005.06.003)] [Medline: [16083778](#)]
13. Xepapadaki P, Papadopoulou NG. Childhood asthma and infection: virus-induced exacerbations as determinants and modifiers. *Eur Respir J* 2010 Aug 31;36(2):438-445 [[FREE Full text](#)] [doi: [10.1183/09031936.00149009](https://doi.org/10.1183/09031936.00149009)] [Medline: [20675781](#)]

14. Puranik S, Forno E, Bush A, Celedón JC. Predicting severe asthma exacerbations in children. *Am J Respir Crit Care Med* 2017 Apr 01;195(7):854-859 [FREE Full text] [doi: [10.1164/rccm.201606-1213PP](https://doi.org/10.1164/rccm.201606-1213PP)] [Medline: [27710010](https://pubmed.ncbi.nlm.nih.gov/27710010/)]
15. Vamathevan J, Clark D, Czodrowski P, Dunham I, Ferran E, Lee G, et al. Applications of machine learning in drug discovery and development. *Nat Rev Drug Discov* 2019 Jun 11;18(6):463-477 [FREE Full text] [doi: [10.1038/s41573-019-0024-5](https://doi.org/10.1038/s41573-019-0024-5)] [Medline: [30976107](https://pubmed.ncbi.nlm.nih.gov/30976107/)]
16. Tison GH, Zhang J, Delling FN, Deo RC. Automated and interpretable patient ECG profiles for disease detection, tracking, and discovery. *Circ Cardiovasc Qual Outcomes* 2019 Sep;12(9):e005289 [FREE Full text] [doi: [10.1161/CIRCOUTCOMES.118.005289](https://doi.org/10.1161/CIRCOUTCOMES.118.005289)] [Medline: [31525078](https://pubmed.ncbi.nlm.nih.gov/31525078/)]
17. Awan SE, Sohail F, Sanfilippo FM, Bennamoun M, Dwivedi G. Machine learning in heart failure: ready for prime time. *Curr Opin Cardiol* 2018 Mar;33(2):190-195. [doi: [10.1097/HCO.0000000000000491](https://doi.org/10.1097/HCO.0000000000000491)] [Medline: [29194052](https://pubmed.ncbi.nlm.nih.gov/29194052/)]
18. Xiong XL, Zhang RX, Bi Y, Zhou WH, Yu Y, Zhu DL. Machine learning models in type 2 diabetes risk prediction: results from a cross-sectional retrospective study in Chinese adults. *Curr Med Sci* 2019 Aug 25;39(4):582-588. [doi: [10.1007/s11596-019-2077-4](https://doi.org/10.1007/s11596-019-2077-4)] [Medline: [31346994](https://pubmed.ncbi.nlm.nih.gov/31346994/)]
19. Spathis D, Vlamos P. Diagnosing asthma and chronic obstructive pulmonary disease with machine learning. *Health Informatics J* 2019 Sep 18;25(3):811-827 [FREE Full text] [doi: [10.1177/1460458217723169](https://doi.org/10.1177/1460458217723169)] [Medline: [28820010](https://pubmed.ncbi.nlm.nih.gov/28820010/)]
20. Goto T, Camargo CA, Faridi MK, Yun BJ, Hasegawa K. Machine learning approaches for predicting disposition of asthma and COPD exacerbations in the ED. *Am J Emerg Med* 2018 Sep;36(9):1650-1654. [doi: [10.1016/j.ajem.2018.06.062](https://doi.org/10.1016/j.ajem.2018.06.062)] [Medline: [29970272](https://pubmed.ncbi.nlm.nih.gov/29970272/)]
21. Exarchos KP, Beltsiou M, Votti CA, Kostikas K. Artificial intelligence techniques in asthma: a systematic review and critical appraisal of the existing literature. *Eur Respir J* 2020 Sep 07;56(3):2000521 [FREE Full text] [doi: [10.1183/13993003.00521-2020](https://doi.org/10.1183/13993003.00521-2020)] [Medline: [32381498](https://pubmed.ncbi.nlm.nih.gov/32381498/)]
22. Patel SJ, Chamberlain DB, Chamberlain JM. A machine learning approach to predicting need for hospitalization for pediatric asthma exacerbation at the time of emergency department triage. *Acad Emerg Med* 2018 Dec 29;25(12):1463-1470 [FREE Full text] [doi: [10.1111/acem.13655](https://doi.org/10.1111/acem.13655)] [Medline: [30382605](https://pubmed.ncbi.nlm.nih.gov/30382605/)]
23. Hosseini A, Buonocore C, Hashemzadeh S, Hojaiji H, Kalantarian H, Sideris C, et al. Feasibility of a secure wireless sensing smartwatch application for the self-management of pediatric asthma. *Sensors (Basel)* 2017 Aug 03;17(8):1780 [FREE Full text] [doi: [10.3390/s17081780](https://doi.org/10.3390/s17081780)] [Medline: [28771168](https://pubmed.ncbi.nlm.nih.gov/28771168/)]
24. Huffaker MF, Carchia M, Harris BU, Kethman WC, Murphy TE, Sakarovich CC, et al. Passive nocturnal physiologic monitoring enables early detection of exacerbations in children with asthma. A proof-of-concept study. *Am J Respir Crit Care Med* 2018 Aug 01;198(3):320-328 [FREE Full text] [doi: [10.1164/rccm.201712-2606OC](https://doi.org/10.1164/rccm.201712-2606OC)] [Medline: [29688023](https://pubmed.ncbi.nlm.nih.gov/29688023/)]
25. Seol HY, Shrestha P, Muth JF, Wi CI, Sohn S, Ryu E, et al. Artificial intelligence-assisted clinical decision support for childhood asthma management: a randomized clinical trial. *PLoS One* 2021;16(8):e0255261 [FREE Full text] [doi: [10.1371/journal.pone.0255261](https://doi.org/10.1371/journal.pone.0255261)] [Medline: [34339438](https://pubmed.ncbi.nlm.nih.gov/34339438/)]
26. Kothalawala DM, Kadalayil L, Weiss VB, Kyyaly MA, Arshad SH, Holloway JW, et al. Prediction models for childhood asthma: a systematic review. *Pediatr Allergy Immunol* 2020 Aug 13;31(6):616-627. [doi: [10.1111/pai.13247](https://doi.org/10.1111/pai.13247)] [Medline: [32181536](https://pubmed.ncbi.nlm.nih.gov/32181536/)]
27. Xiong S, Chen W, Jia X, Jia Y, Liu C. Machine learning for prediction of asthma exacerbations among asthmatic patients: a systematic review and meta-analysis. *BMC Pulm Med* 2023 Jul 28;23(1):278 [FREE Full text] [doi: [10.1186/s12890-023-02570-w](https://doi.org/10.1186/s12890-023-02570-w)] [Medline: [37507662](https://pubmed.ncbi.nlm.nih.gov/37507662/)]
28. McGowan J, Sampson M, Salzwedel DM, Cogo E, Foerster V, Lefebvre C. PRESS peer review of electronic search strategies: 2015 guideline statement. *J Clin Epidemiol* 2016 Jul;75:40-46 [FREE Full text] [doi: [10.1016/j.jclinepi.2016.01.021](https://doi.org/10.1016/j.jclinepi.2016.01.021)] [Medline: [27005575](https://pubmed.ncbi.nlm.nih.gov/27005575/)]
29. The world's #1 systematic review tool. Covidence. URL: <https://www.covidence.org/> [accessed 2024-04-29]
30. AlSaad R, Malluhi Q, Janahi I, Boughorbel S. Predicting emergency department utilization among children with asthma using deep learning models. *Healthcare Anal* 2022 Nov;2:100050 [FREE Full text] [doi: [10.1016/j.health.2022.100050](https://doi.org/10.1016/j.health.2022.100050)]
31. Bhardwaj P, Tyagi A, Tyagi S, Antão J, Deng Q. Machine learning model for classification of predominantly allergic and non-allergic asthma among preschool children with asthma hospitalization. *J Asthma* 2023 Mar;60(3):487-495 [FREE Full text] [doi: [10.1080/02770903.2022.2059763](https://doi.org/10.1080/02770903.2022.2059763)] [Medline: [35344453](https://pubmed.ncbi.nlm.nih.gov/35344453/)]
32. Bose S, Kenyon CC, Masino AJ. Personalized prediction of early childhood asthma persistence: a machine learning approach. *PLoS One* 2021;16(3):e0247784 [FREE Full text] [doi: [10.1371/journal.pone.0247784](https://doi.org/10.1371/journal.pone.0247784)] [Medline: [33647071](https://pubmed.ncbi.nlm.nih.gov/33647071/)]
33. Deliu M, Fontanella S, Haider S, Sperrin M, Geifman N, Murray C, et al. Longitudinal trajectories of severe wheeze exacerbations from infancy to school age and their association with early-life risk factors and late asthma outcomes. *Clin Exp Allergy* 2020 Mar;50(3):315-324 [FREE Full text] [doi: [10.1111/cea.13553](https://doi.org/10.1111/cea.13553)] [Medline: [31876035](https://pubmed.ncbi.nlm.nih.gov/31876035/)]
34. Deng X, Thurston G, Zhang W, Ryan I, Jiang C, Khwaja H, et al. Application of data science methods to identify school and home risk factors for asthma and allergy-related symptoms among children in New York. *Sci Total Environ* 2021 May 20;770:144746 [FREE Full text] [doi: [10.1016/j.scitotenv.2020.144746](https://doi.org/10.1016/j.scitotenv.2020.144746)] [Medline: [33736384](https://pubmed.ncbi.nlm.nih.gov/33736384/)]
35. Gorham TJ, Tumin D, Groner J, Allen E, Retzke J, Hersey S, et al. Predicting emergency department visits among children with asthma in two academic medical systems. *J Asthma* 2023 Dec;60(12):2137-2144 [FREE Full text] [doi: [10.1080/02770903.2023.2225603](https://doi.org/10.1080/02770903.2023.2225603)] [Medline: [37318283](https://pubmed.ncbi.nlm.nih.gov/37318283/)]

36. Habukawa C, Ohgami N, Matsumoto N, Hashino K, Asai K, Sato T, et al. A wheeze recognition algorithm for practical implementation in children. *PLoS One* 2020 Oct 8;15(10):e0240048 [FREE Full text] [doi: [10.1371/journal.pone.0240048](https://doi.org/10.1371/journal.pone.0240048)] [Medline: [33031408](https://pubmed.ncbi.nlm.nih.gov/33031408/)]
37. Hee HI, Balamurali BT, Karunakaran A, Herremans D, Teoh OH, Lee KP, et al. Development of machine learning for asthmatic and healthy voluntary cough sounds: a proof of concept study. *Appl Sci* 2019 Jul 16;9(14):2833. [doi: [10.3390/app9142833](https://doi.org/10.3390/app9142833)]
38. Hogan AH, Brimacombe M, Mosha M, Flores G. Comparing artificial intelligence and traditional methods to identify factors associated with pediatric asthma readmission. *Acad Pediatr* 2022;22(1):55-61 [FREE Full text] [doi: [10.1016/j.acap.2021.07.015](https://doi.org/10.1016/j.acap.2021.07.015)] [Medline: [34329757](https://pubmed.ncbi.nlm.nih.gov/34329757/)]
39. Hurst JH, Zhao C, Hostetler HP, Ghiasi Gorveh M, Lang JE, Goldstein BA. Environmental and clinical data utility in pediatric asthma exacerbation risk prediction models. *BMC Med Inform Decis Mak* 2022 Apr 22;22(1):108 [FREE Full text] [doi: [10.1186/s12911-022-01847-0](https://doi.org/10.1186/s12911-022-01847-0)] [Medline: [35459216](https://pubmed.ncbi.nlm.nih.gov/35459216/)]
40. Krautenbacher N, Flach N, Böck A, Laubhahn K, Laimighofer M, Theis FJ, et al. A strategy for high-dimensional multivariable analysis classifies childhood asthma phenotypes from genetic, immunological, and environmental factors. *Allergy* 2019 Jul;74(7):1364-1373 [FREE Full text] [doi: [10.1111/all.13745](https://doi.org/10.1111/all.13745)] [Medline: [30737985](https://pubmed.ncbi.nlm.nih.gov/30737985/)]
41. Messinger AI, Bui N, Wagner BD, Szeffler SJ, Vu T, Deterding RR. Novel pediatric-automated respiratory score using physiologic data and machine learning in asthma. *Pediatr Pulmonol* 2019 Aug;54(8):1149-1155 [FREE Full text] [doi: [10.1002/ppul.24342](https://doi.org/10.1002/ppul.24342)] [Medline: [31006993](https://pubmed.ncbi.nlm.nih.gov/31006993/)]
42. Seol HY, Rolfes MC, Chung W, Sohn S, Ryu E, Park MA, et al. Expert artificial intelligence-based natural language processing characterises childhood asthma. *BMJ Open Respir Res* 2020 Feb;7(1):e000524 [FREE Full text] [doi: [10.1136/bmjresp-2019-000524](https://doi.org/10.1136/bmjresp-2019-000524)] [Medline: [33371009](https://pubmed.ncbi.nlm.nih.gov/33371009/)]
43. Sills MR, Ozkaynak M, Jang H. Predicting hospitalization of pediatric asthma patients in emergency departments using machine learning. *Int J Med Inform* 2021 Jul;151:104468 [FREE Full text] [doi: [10.1016/j.ijmedinf.2021.104468](https://doi.org/10.1016/j.ijmedinf.2021.104468)] [Medline: [33940479](https://pubmed.ncbi.nlm.nih.gov/33940479/)]
44. Wolff RF, Moons KG, Riley RD, Whiting PF, Westwood M, Collins GS, PROBAST Group†. PROBAST: a tool to assess the risk of bias and applicability of prediction model studies. *Ann Intern Med* 2019 Jan 01;170(1):51-58 [FREE Full text] [doi: [10.7326/M18-1376](https://doi.org/10.7326/M18-1376)] [Medline: [30596875](https://pubmed.ncbi.nlm.nih.gov/30596875/)]
45. Luo W, Phung D, Tran T, Gupta S, Rana S, Karmakar C, et al. Guidelines for developing and reporting machine learning predictive models in biomedical research: a multidisciplinary view. *J Med Internet Res* 2016 Dec 16;18(12):e323 [FREE Full text] [doi: [10.2196/jmir.5870](https://doi.org/10.2196/jmir.5870)] [Medline: [27986644](https://pubmed.ncbi.nlm.nih.gov/27986644/)]
46. Jeddi Z, Gryech I, Ghogho M, El Hammoumi M, Mahraoui C. Machine learning for predicting the risk for childhood asthma using prenatal, perinatal, postnatal and environmental factors. *Healthcare (Basel)* 2021 Oct 29;9(11):1464 [FREE Full text] [doi: [10.3390/healthcare9111464](https://doi.org/10.3390/healthcare9111464)] [Medline: [34828510](https://pubmed.ncbi.nlm.nih.gov/34828510/)]
47. Ferri P, Romero-Garcia N, Badenes R, Lora-Pablos D, Morales TG, Gómez de la Cámara A, et al. Extremely missing numerical data in electronic health records for machine learning can be managed through simple imputation methods considering informative missingness: a comparative of solutions in a COVID-19 mortality case study. *Comput Methods Programs Biomed* 2023 Dec;242:107803 [FREE Full text] [doi: [10.1016/j.cmpb.2023.107803](https://doi.org/10.1016/j.cmpb.2023.107803)] [Medline: [37703700](https://pubmed.ncbi.nlm.nih.gov/37703700/)]
48. Li J, Yan XS, Chaudhary D, Avula V, Mudiganti S, Husby H, et al. Imputation of missing values for electronic health record laboratory data. *NPJ Digit Med* 2021 Oct 11;4(1):147 [FREE Full text] [doi: [10.1038/s41746-021-00518-0](https://doi.org/10.1038/s41746-021-00518-0)] [Medline: [34635760](https://pubmed.ncbi.nlm.nih.gov/34635760/)]
49. Elshawi R, Al-Mallah MH, Sakr S. On the interpretability of machine learning-based model for predicting hypertension. *BMC Med Inform Decis Mak* 2019 Jul 29;19(1):146 [FREE Full text] [doi: [10.1186/s12911-019-0874-0](https://doi.org/10.1186/s12911-019-0874-0)] [Medline: [31357998](https://pubmed.ncbi.nlm.nih.gov/31357998/)]
50. Zheng L, Ohde JW, Overgaard SM, Brereton TA, Jose K, Wi C, et al. Clinical needs assessment of a machine learning-based asthma management tool: user-centered design approach. *JMIR Form Res* 2024 Jan 15;8:e45391 [FREE Full text] [doi: [10.2196/45391](https://doi.org/10.2196/45391)] [Medline: [38224482](https://pubmed.ncbi.nlm.nih.gov/38224482/)]

Abbreviations

AI: artificial intelligence

ANN: artificial neural network

AUROC: area under the receiver operating characteristic curve

ED: emergency department

EHR: electronic health record

ML: machine learning

NLP: natural language processing

NPV: negative predictive value

PRISMA-ScR: Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews

PROBAST: Prediction Model Risk of Bias Assessment Tool

RF: random forest

SVM: support vector machine

Edited by K El Emam, B Malin; submitted 01.03.24; peer-reviewed by M van der Kamp, HL Ooi; comments to author 07.04.24; revised version received 27.05.24; accepted 13.06.24; published 27.08.24.

Please cite as:

*Ojha T, Patel A, Sivapragasam K, Sharma R, Vosoughi T, Skidmore B, Pinto AD, Hosseini B
Exploring Machine Learning Applications in Pediatric Asthma Management: Scoping Review
JMIR AI 2024;3:e57983*

URL: <https://ai.jmir.org/2024/1/e57983>

doi: [10.2196/57983](https://doi.org/10.2196/57983)

PMID: [39190449](https://pubmed.ncbi.nlm.nih.gov/39190449/)

©Tanvi Ojha, Atushi Patel, Krishihan Sivapragasam, Radha Sharma, Tina Vosoughi, Becky Skidmore, Andrew D Pinto, Banafshe Hosseini. Originally published in JMIR AI (<https://ai.jmir.org>), 27.08.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR AI, is properly cited. The complete bibliographic information, a link to the original publication on <https://www.ai.jmir.org/>, as well as this copyright and license information must be included.

Viewpoint

Regulatory Frameworks for AI-Enabled Medical Device Software in China: Comparative Analysis and Review of Implications for Global Manufacturer

Yu Han¹; Aaron Ceross¹; Jeroen Bergmann^{1,2}

¹University of Oxford, Oxford, United Kingdom

²Department of Technology and Innovation, The University of Southern Denmark, Denmark, Denmark

Corresponding Author:

Yu Han

University of Oxford

Old Road Campus

Oxford, OX1 2JD

United Kingdom

Phone: 44 789314203

Email: yu.han@eng.ox.ac.uk

Abstract

The China State Council released the new generation artificial intelligence (AI) development plan, outlining China's ambitious aspiration to assume global leadership in AI by the year 2030. This initiative underscores the extensive applicability of AI across diverse domains, including manufacturing, law, and medicine. With China establishing itself as a major producer and consumer of medical devices, there has been a notable increase in software registrations. This study aims to study the proliferation of health care-related software development within China. This work presents an overview of the Chinese regulatory framework for medical device software. The analysis covers both software as a medical device and software in a medical device. A comparative approach is employed to examine the regulations governing medical devices with AI and machine learning in China, the United States, and Europe. The study highlights the significant proliferation of health care-related software development within China, which has led to an increased demand for comprehensive regulatory guidance, particularly for international manufacturers. The comparative analysis reveals distinct regulatory frameworks and requirements across the three regions. This paper provides a useful outline of the current state of regulations for medical software in China and identifies the regulatory challenges posed by the rapid advancements in AI and machine learning technologies. Understanding these challenges is crucial for international manufacturers and stakeholders aiming to navigate the complex regulatory landscape.

(JMIR AI 2024;3:e46871) doi:[10.2196/46871](https://doi.org/10.2196/46871)

KEYWORDS

NMPA; medical device software; device registration; registration pathway; artificial intelligence; machine learning; medical device; device development; China; regulations; medical software

Background

New software solutions that are being developed, especially medical devices that combine artificial intelligence (AI) and machine learning (ML), show a huge potential for patient benefit. These kinds of applications can be used across different medical conditions, with the potential for easy scale-up to larger populations. It can reduce the burden on health care professionals and decrease the possible risk of missing vital information. For example, radiology software is used to screen and diagnose large amounts of X-ray images [1]. A combined AI and ML approach can also be applied in, for example, oncology for the next-generation sequencing [2], in

ophthalmology for image recognition [3], or as a support system for general medical decision-making [4]. ML models have been used for anything from improving outcomes for diabetic patients [5] to tuberculosis diagnosis [6]. Many of these approaches should be applicable on a global scale, and thus there is a growing interest in applying these solutions across borders. This has led clinicians, academics, and manufacturers to look at China and its medical device regulatory environment. However, navigating China's regulatory environment presents inherent complexities stemming from language barriers, geographical distances, and a general lack of familiarity with the regulatory framework. These complexities are augmented by innovative products that can have unconventional regulatory requirements.

Easing these barriers holds the potential to facilitate the seamless exchange of solutions across international boundaries, fostering mutual opportunities. This paper provides a regulatory view of China, the biggest booming market for medical device software, and discusses the implications for global manufacturers.

China AI Development Plan

The 21st century has seen a rapid development of the Chinese economy and its ability to produce, manufacture, and distribute technology. In 2017, the China State Council published a white paper discussing a *new generation AI development plan* [7]. The document indicated that the number of AI scientific papers published and invention patents granted in China ranked second worldwide. Several domain-specific applications that were developed in China have gained widespread attention, including intelligent monitoring, biometric recognition, industrial robots, service robots, and unmanned driving. The AI Development Plan clearly states China's support for smart medical care, products, and services that use AI. Moreover, it is stated that this even should be developed as a priority. The vision is to establish a major medical system that leverages AI and ML.

China has become a major global producer and consumer of medical devices [8]. With one of the world's largest populations (1.426 billion in 2022) [9], the need is obvious in terms of access to medical technology. In 2019, the Chinese medical device market had an estimated revenue value of 629 billion RMB (US \$88.7 billion), more than double of what it was in 2015 (308 billion RMB or US \$44.2 billion) before the plan was released [10]. This coincides with a growing trend of medical device software (MDSW) registrations [11]. One factor driving this trend is the potential that digital health offers in terms of ease of scalability, which provides an opportunity to advance health care more sustainably.

Table 1. China National Medical Products administration (NMPA) regulatory documents for medical device software.

Date of publication	Regulatory document
August 2015	Guidelines of medical device software registration and review [12]
July 2019	Key points of deep learning decision-making assisting medical device software review [13]
July 2021	Guidelines for the classification and designation of artificial intelligence medical software [14]
March 2022	Guidelines of medical device software registration and review [15]
August 2022	Guidelines for the classification and designation of artificial intelligence medical software [16]

Several standards are referenced in the regulation, and they include (but are not limited to) standards on the risk level of software (YY/T0664-2008), on software engineering (GB/T 19003-2008), and those that describe the medical device quality management requirements (YY/T 0287-2003). These standards can help with compliance with these new regulations, and this provides a useful function in the regulatory pathway.

In China, MDSW includes “software as a medical device” (SaMD) and “software in a medical device” (SiMD). The term “software as a medical device” is defined by the International Medical Device Regulators Forum (IMDRF) as software intended to be used for one or more medical purposes without being part of a hardware medical device [17]. This delineation

Global manufacturers seeking to enter the Chinese market must possess a profound comprehension of the regulatory landscape governing MDSW. This necessitates a thorough grasp of the intricacies surrounding registration prerequisites, regulatory oversight, disparities vis-à-vis regulatory bodies in alternative geographic regions, and the contemporary device taxonomy specific to China. Simultaneously, researchers and health care practitioners must remain vigilant by staying abreast of the latest developments transpiring within the Chinese milieu. The global pandemic has unequivocally underscored the imperative of comprehending and navigating policies and regulations in foreign jurisdictions, including but not limited to China, as an indispensable facet of effectively addressing worldwide crises. By extension, software-based solutions can similarly accrue significant advantages through adopting a holistic and globally informed perspective.

Chinese Regulation on Medical Device Software

After the *new generation AI development plan* was introduced, China's medical products regulatory authority—National Medical Products Administration (NMPA)—released many regulations to fit the plan's theme. In 2022, the NMPA launched a program on digital health. Two MDSW guidelines were published as part of this program. Table 1 shows a series of regulatory documents published with regard to MDSW and AI-enabled software. The NMPA released the first document in 2015, while a more up-to-date document was made public in 2022. This updated version raised more detailed requirements for the whole life cycle management of these technologies, as well as for quality management, verification, raw code analysis, and safety management.

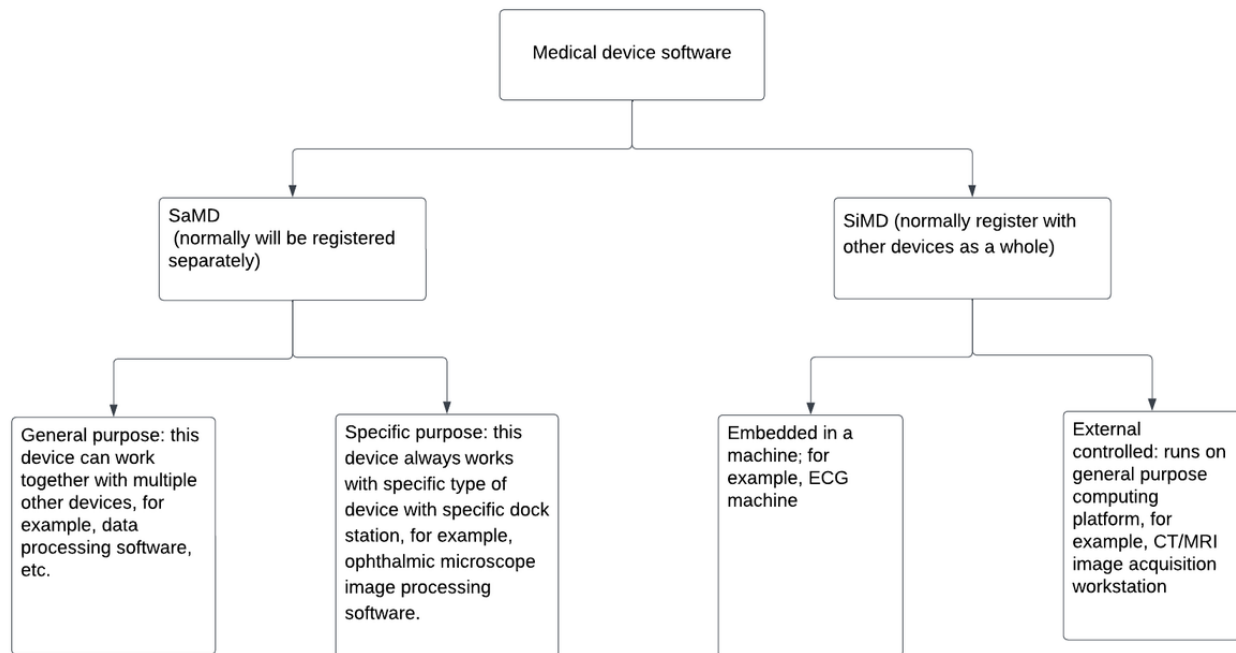
posits the software itself as a standalone medical device. Conversely, “SiMD” denotes software that functions as an integral constituent of an entire medical apparatus, such as its involvement in the operation of magnetic resonance imaging scanners, x-ray machines, or insulin pumps. In these cases, the software and other components all fall under the same registration license “SiMD.” It is noteworthy that in China, software harnessing AI or ML technologies may concurrently straddle both the SaMD and SiMD categories.

An overview is given in Figure 1 with regard to how software devices are categorized from a function or a design perspective. Devices are initially split into SaMD and SiMD. SaMD is normally registered separately, while, as mentioned previously

in the case of SiMD, the software is often registered along with other components [15]. In the case of SiMD, the software doesn't have its own classification, but it shares the same classification with other parts of the device. The final classification would then be based on the risk of the whole device. SaMD can be split into 2 types depending on its purposes. Its purpose can be (1) general or (2) specific. For the general-purpose definition, the device can work together with multiple other devices, as happens in the example of data processing software. For the specific purpose case, the device

always works with a distinct set of devices for a particular purpose. An illustration of this is the ophthalmic microscope image processing software. The SiMD also consists of 2 types of devices. One type is embedded in a machine (eg, an electrocardiogram machine), while the other type is externally controlled. A general-purpose computing platform (eg, a computed tomography [CT] and magnetic resonance image acquisition workstation) is a good exemplification of an externally controlled type of SiMD. The categorization of the software is a crucial step in the regulatory journey of a product.

Figure 1. Categories of medical device software. CT: computed tomography; ECG: electrocardiography; MRI: magnetic resonance imaging; SaMD: software as a medical device; SiMD: software in a medical device.



Regulatory Environment in China

The oversight and governance of medical devices within China are primarily administered by the Center for Medical Device Evaluation, an integral component of the NMPA. The regulatory landscape formulated by the NMPA to govern medical devices is predicated upon a comprehensive framework rooted in Chinese legislation, regulations, and advisory directives. This multifaceted regulatory apparatus encapsulates various facets pertinent to market entry, encompassing the specification of device categories, the classification of devices, the requisite content of registration review dossiers, and the imperative facet of post-market surveillance. In conformity with these regulatory imperatives, manufacturers need to engage proactively with the NMPA, necessitating their involvement across all aforementioned dimensions.

Medical devices are subject to regulatory oversight within a risk management framework that stratifies these products according to risk levels, ranging from low risk (class I) to high risk (class III). In the case of manufacturers engaging in the importation of medical devices into China, the responsibility for the review process falls under the purview of national authorities. Concurrently, certain domestically produced medical

devices are subject to regulatory scrutiny by provincial authorities. The classification of a medical device within the Chinese regulatory context necessitates the alignment of its device description with the pertinent information contained within the medical device catalog [18].

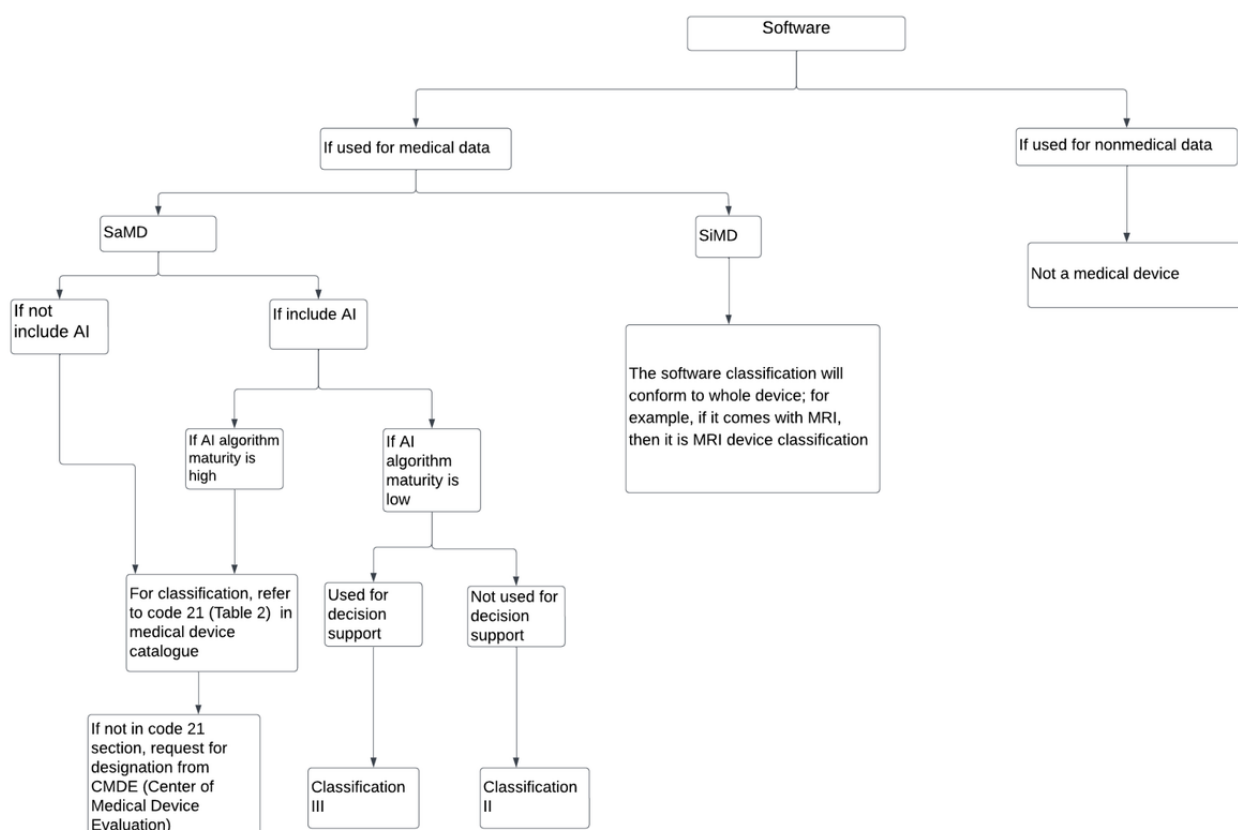
In general, manufacturers possess 2 principal avenues for conceiving innovative medical equipment, which are occasionally amenable to synergistic integration. The first approach involves commencing with a patient-centered needs assessment (need-led innovation) to engender a “novel” technological solution. The alternative approach entails the development of a “novel” technology, with the subsequent identification of a correlating patient need [19]. These innovations can occur either before or after appropriate regulations have been set [20]. It is common that transformative ideas initially do not have suitable regulations in place and that this mismatch can lead to either delays in market adoption or concerns in terms of device performance and safety. However, any medical software enterprise aspiring to introduce its product to the market is mandated to adhere to prevailing regulatory mandates. Accordingly, a comprehensive comprehension of the product's classification and regulatory prerequisites within a specific market is of paramount significance, as the realms of innovation and regulation engage in a dynamic interplay. A

good understanding is particularly important, as it has been suggested that the complexity of medical device regulations can increase whenever new regulations are formed [21]. Erroneous classification of product risk and the correlated regulatory obligations can result in exacerbated time and financial investments for subsequent rectification. Thus, the incorporation of regulatory considerations should be undertaken expeditiously, as many decisions regarding the final product are already made at the early stages of the research and development process.

Specific Rules for Software and AI

Medical software is basically divided into auxiliary diagnosis and treatment devices according to their intended use. A detailed translation of the software catalog can be found in [Multimedia Appendix 1](#). The SaMD (which begins with code 21 according to regulation) is categorized into 6 categories: treatment planning software (21-01), image processing software (21-02), data processing software (21-03), decision support software (21-04), in vitro diagnostic software (21-05), and other software (21-06). If the device to be registered is not included in the list, then it has to be re-classified through the device designation pathway [22]. A simple flowchart for the classification of the software is shown in [Figure 2](#).

Figure 2. Medical device software classification flowchart. AI: artificial intelligence; MRI: magnetic resonance imaging; SaMD: software as a medical device; SiMD: software in a medical device.



There are 2 branches for SaMD, which are split between AI and other technologies. If AI is applied, then a further decision is made according to the level of maturity of the algorithm. A high maturity level of the algorithm signifies that the safety and efficacy profiles of the algorithm have been judiciously established, while conversely, a lower degree of maturity implies that such establishment has not been ascertained. A preconsultation meeting could be used to discuss the maturity level with the NMPA. If the AI algorithm has a well-established profile, then manufacturers can refer back to [Multimedia Appendix 1](#), code 21 [18] for classification. A request for designation could then be sent to the NMPA, if the device is out of scope. If the maturity degree is low, then there are 2 classifications possible. The device could be classified as a class

III device if it is used for decision support; otherwise, the device will be assigned a classification of II, which represents a lower risk class. According to the Medical Device Classification Catalog [18], a class II device classification is given when the software does not contain any AI and the medical software is only used for image and data processing, thus not used for diagnostics. If it were used for diagnostic purposes, then the classification would become III. The degree of risk for diagnostic software is determined by the level of maturity, registration of the applied algorithm in their database, and the “object” of interest (referring to a particular disease, such as a certain type of cancer) [23].

However, if the software just provides diagnostic suggestions through its algorithm (in other words, it only has an auxiliary

diagnostic function and does not directly provide a diagnostic conclusion), then the device can be regulated as a class II medical device. Yet, if the diagnostic software automatically recognizes, for instance, a lesion site through its algorithm and provides clear diagnostic prompts, a class III classification would be assigned due to the increased relative risk. In general, medical software using AI technology is currently managed by designating it the highest possible classification in China. This is driven by the novelty of the technology, as well as the lack of in-depth and complete evaluations of the clinical risks. China has been focusing more on reviewing the algorithm itself, while in the United States, attention has shifted toward the manufacturers themselves [24].

It should be noted that not all software applied in the medical field is regulated as a medical device by the NMPA. If the software is used to process medical device data for measurement, model calculation, or analysis, then it is deemed MDSW and thus regulated by NMPA. If the software is used for non-medical device data, it will not be regulated as a medical

device under the NMPA. This is the case when software is used for the processing of general patient information or for patient testing reports, both of which are not seen as medical device data.

China's and the IMDRF criteria share many similarities on how to determine if the software is a SaMD. According to the IMDRF [25], the SaMD definition should include a clear statement about the intended use of the device, and the following aspects need to be described in order to be able to be regulated as SaMD (Textbox 1).

In alignment with the IMDRF, the European Medicines Agency declares that only devices whose intended use includes a medical purpose and influences the patient's health care situation can be deemed to be medical devices. Products such as medical information management software (which is a hospital management tool) are also not designated as medical devices. This is similar to China, since it then does not meet the definition of a medical device.

Textbox 1. Aspects need to be described in order to be able to be regulated as software as a medical device (SaMD).

- The “significance of the information provided by the SaMD to the health care decision,” which is used to identify the intended medical purpose of the SaMD.
- The “state of the health care situation or condition” that the SaMD is intended for.

General Registration Process and Clinical Evaluation

Ordinarily, medical software devices, regardless of whether they use AI or not, are typically not categorized under class I. Within the context of classes II and III devices, the registration process typically takes around 18 months if no clinical trials are required. However, once clinical trials are needed, the registration timeline can extend to around 36 months or sometimes even longer. The exact timeline is dependent on the complexity of the device and the associated clinical data. It is of particular importance to note that certain devices may qualify for expedited processing through a fast-track pathway. Under these circumstances, not only can registration fees be exempted, but the registration timeline is accelerated, as it is typically condensed to approximately 50 working days. Currently, there are 2 software devices that have been designated under the Fast Track pathway in China, namely, an implantable left ventricular assist software system and a coronary CT fractional flow reserve calculation software, identified by license numbers 20213120987 and 20213210270, respectively. Refer to Figure 3 for an overview of the registration process for Class II and Class III software devices.

In accordance with the Notice of the Chinese NMPA, which relates to the issuance of 5 technical guidelines, including the Technical Guidelines for Clinical Evaluation of Medical Devices (number 73 of 2021), it is evident that there exist 3 distinct pathways for meeting the required clinical standards (see Figure 3). These pathways encompass (1) a clinical exemption, (2) a

clinical comparison, and (3) a clinical trial, each associated with a gradient of clinical requirements ranging from low to high. Exemptions can be obtained if the device is part of the catalog of devices that are exempt. For medical devices not encompassed within the “Catalog of Medical Devices Exempt from Clinical Trials,” the pathway of conducting a comparative analysis with similar products already available on the market can be explored. This can be realized through the systematic collection and meticulous analysis of clinical data and other pertinent evidence, thereby proofing their equivalence and thus expediting the clinical evaluation process.

The need to conduct clinical trials for AI medical devices is thus not universally mandated. Furthermore, if clinical trials are required, then it is not determined solely by their classification. The requirement to run a clinical trial depends on the intent and the application. The NMPA's “Guidelines for the Evaluation of Artificial Intelligence Medical Devices,” states that for functionalities that do not entail decision-making assistance and are grounded in core operations, a rigorous comparative analysis with similar medical devices within the same category is required. However, for decision-assistive functions underpinned by core algorithms, a comparative analysis with equivalent medical devices within the same category is only advocated. Nonetheless, the devices selected for comparison should ideally have undergone comprehensive clinical trials, although historical data may be acceptable in certain circumstances. Finally, novel functions, algorithms, and applications should be subjected to exhaustive clinical trials to ensure their efficacy and safety within the clinical domain.

Figure 3. Medical device software registration process. NMPA: National Medical Products Administration.



Cultivating AI Software Devices: An Emerging Trend in China

Following the introduction of the new generation AI development plan, major shifts have occurred in both investment and policy domains to align with the overarching objectives of this plan. Notably, the NMPA, as China's regulatory authority for medical products, has promulgated a series of regulations in line with the thematic contours of the plan. In 2022, the NMPA initiated a digital health program. Over the course of 5 years, the NMPA, operating as a subsidiary of the Chinese government, has enacted a suite of regulations to govern the medical device industry, a selection of which is delineated in Table 1. These encompass pivotal documents such as the “Key Points of Deep Learning Decision-Making Assisting Medical

Device Software Review [13]” and the “Guidelines for the Classification and Designation of Artificial Intelligence Medical Software [14],” as well as the “Guidelines for Medical Device Software Registration and Review [15]” and a duplicate mention of the “Guidelines for the Classification and Designation of Artificial Intelligence Medical Software” [16].

China's concerted efforts in this domain have manifested in substantial investments and the development of numerous medical software applications. An illustrative milestone occurred in the year 2020 when the first AI-based diagnostic software received approval in China, specifically for employment in CT image AI-assisted diagnostic software products. As of 2023, an exhaustive review of the NMPA website has revealed that China has granted approvals for more than 50 AI medical device products rooted in deep learning technology [26]. These

products, predominantly classified as medical software, serve as pivotal aids in diagnostic processes encompassing CT images, fundus images, and magnetic resonance images, and are strategically deployed within specialized fields such as radiology, ophthalmology, and cardiology. Moreover, regional governments seem to have demonstrated proactive engagement with the evolving landscape.

Challenges Posed by Software and AI in Medical Devices

As AI technology develops further, regulators will also face the challenge of applying regulatory safeguards to these novel technologies. The technical complexity of certain medical software solutions warrants the description of these systems as a “black box,” due to their inherent opacity [27,28]. In addition, traditional frameworks for regulation are not suitable for adaptive AI and ML technologies, since the algorithms are constantly learning and making changes [29]. Therefore, digital health care solutions provide a different set of challenges to regulators and the traditional fixed regulatory framework is not suitable for this type of AI device. At present, governmental agencies in the United States, the European Union, and China have all issued new regulatory methods or frameworks for MDSW to help cope with the changing landscape.

The regulation of AI devices is to ensure safety, quality, and reliability requirements are met. One key concern is the “unlocked” nature of these devices. “Locked” devices mean that the algorithm provides the exact same result for a (specific) given input [29]. This contrasts with an “unlocked” algorithm, which represents a continuous learning algorithm. The “unlocked” algorithm is also known as an adaptive algorithm, and it changes its behavior using a predefined learning process that provides time-based updates from new data with the overall aim of improving its clinical performance. This algorithm continuously changes the input-output relationship. Thus, for a given set of inputs, the output may be different before and after these changes are implemented. This means that after a “locked” device has been approved and given access to the market, the device can continue to self-learn and thus alter its performance in comparison to when it was first approved. In this situation, it is difficult for the clinicians or the authorities to fully trust the device before they use it in practice. So far, the Food and Drug Administration (FDA) has not yet approved a device that integrates continual learning AI, as they have only granted approval to locked systems [29].

The FDA has enacted the Digital Health Innovation Action Plan [30], with the aim of building a more dynamic approval process with precertification for companies that will then have the ability to change the characteristics of a product without needing ongoing FDA assessment. This enterprise-based approach (precertification program) is very different from traditional medical device regulation. The FDA adopted the precertification program together with the total product life cycle database to screen for eligible organizations. They also adopted a “predetermined change control plan.” This plan provides a complete approach based on the total product lifecycle in a way that manages the risk to patients in a controlled manner.

The European Union (EU) also enacted new directives to regulate this fast-changing technology domain. They include the general data protection regulation (GDPR), cybersecurity directive, medical devices regulation, and in vitro diagnostic medical device regulation. The GDPR and the Cybersecurity Directive took effect in May 2018, whilst the medical devices regulation was applied in May 2021, with the in vitro diagnostic medical device regulation following suit a year later. These recent changes further highlight the moving landscape of regulations on a global scale.

Besides the apprehension about the increase in regulatory complexity for AI and ML, other aspects are also starting to raise concerns. Among those are ethical considerations, cybersecurity, and the reproducibility of the performance. These aspects are briefly discussed below.

Ethical Considerations

Ethical issues have been intensely debated since the start of AI technology development. In the medical field, obvious questions are posed with regard to data privacy, physician dependency, and potential bias in post-GDPR algorithms, as well as concerns about changes in the doctor-patient relationship [31]. People are also concerned about algorithmic fairness and potential biases. The algorithms are data-driven, and it could be that the data used might not meet the required ethical standards.

In April 2019, the National Artificial Intelligence Standardization General Group in China issued the “Artificial Intelligence Ethical Risk Analysis Report” [32], which further clarified that the principle of fundamental human interests should be considered from three viewpoints: (1) the impact on society, (2) the AI algorithm, and (3) the used data. All these ethical concerns need to be navigated in order to create appropriate technology that can be used in the clinic.

Reproducibility

Reproducibility is also an important item in the field of AI. Nowadays, many AI devices face a problem as their outcomes are not verifiable by third parties [33]. The reasons for this can be related to the quality of the data, data inputs, the transparency of data, or the code used for processing, to name a few factors [34]. There is a particular concern for adaptive AI, as the data upon which the model would be built changes, which in turn can trigger a change in outputs. Consideration should also be given to the need for detailed information on the data processing and training pipelines, as this is often lacking [35].

NMPA issued a document (number 8, 2022) [16] that requires reproducibility evidence from the sponsor in multiple dossier sections. These sections include user need analysis, algorithm property evaluation, and algorithm verification and validation. In the algorithm property evaluation, it suggests that applicants should consider requirements such as false negatives and false positives (indicators and relationships), repeatability, reproducibility, and robustness. At the same time, all factors that affect algorithm performance should be analyzed, and their degree of influence should be determined. This includes things such as the acquisition equipment, acquisition parameters,

disease composition, and lesion characteristics, among others. Taking these into account will improve algorithm interpretability and it can serve as the basis for software verification and validation [36].

Cybersecurity

Like other computer systems, MDSW can be vulnerable to security breaches [37]. It has been suggested that 53% of connected medical devices contain critical vulnerabilities, and health care professionals struggle to maintain the inventories of connected devices [38]. For many years, cyberattacks have been identified as the top health tech hazard within this space [38]. The FDA indicates that cybersecurity issues could directly impact the safety and effectiveness of the device, as further harm can be caused to the patients who are using them [37]. Reducing cybersecurity risks is especially challenging while medical devices interact with human bodies; as a result, it becomes a multidisciplinary problem concerning engineering, computer science, medical, and physical sciences.

The IMDRF issued Principles and Practices for Medical Device Cybersecurity in 2020 [39], which introduces a “total product life cycle” risk reduction plan for cybersecurity. Authorities are now focusing on scrutinizing applicants’ dossiers to make sure a thorough plan has been designed, which contains a risk management process, risk analysis, risk control or residual risk, post-marketing plan, etc. In 2022, the NMPA released a new version of principles of medical device cybersecurity technical evaluation [40], which also ensures data confidentiality, integrity, availability, authenticity, accountability, nonrepudiation, and reliability are covered according to GB/t 29246-2012. The NMPA suggests that applicants make sure that the risk management method is applied throughout the whole life cycle to ensure patient safety. They will focus on quality control across all stages mentioned before in both the pre- and postmarket phases.

Future Directions

Since China joined the IMDRF in 2013 [41], China has adopted and referenced international regulatory methods when formulating its own regulations. Regulatory similarities between China and other countries have been witnessed and demonstrated. However, China also has its own local requirements, standards, and regulatory ideologies, which can be an additional layer of complexity for global manufacturers who want to bring their medical devices to the Chinese market.

There are different aspects for global manufacturers to pay attention to when they want to leverage US or EU experience for the Chinese market. In China, the focus is more on the maturity of the algorithm, which is different from the FDA sponsor qualification program. Differences in sample populations upon which the algorithm is built are another key consideration, in addition to the requirement to ensure data confidentiality and the protection of patients in a specific region. In the Regulatory Science Action Plan issued by NMPA in 2019

[42], there is a clear focus on AI, which suggests more regulations might be developed with an increased level of harmonization with the US, EU, or other markets. Nonetheless, regulatory inconsistency still exists between countries. The same device can be regulated very differently across borders, which poses global manufacturers with big challenges. Large, well-founded medical device companies usually have global regulatory affairs professionals that deal with this situation, but innovation may also arise from small research teams at universities or innovative small and medium enterprises. In this situation, the complexity of the regulatory environment will hinder the potential of influential new products to enter the market. The regulatory strategy will need to differ from region to region to provide the best possible match for each.

For example, in the United States, a high-tech device could be registered as a class II device if it is like a predicate device that has already been registered. In this case, the characteristics need to be the same, and there should not be any cause for concern with regard to the safety and effectiveness of the device. However, in China, manufacturers will need to refer to the classification catalog, which aims to classify the device based on its own safety and effectiveness. If it is a high-tech device, then it becomes more likely that it will be seen as a class III device in China. This means that the device will face more stringent registration requirements, including clinical evaluation and even trials. Manufacturers need to consider this when they start to map their market potential globally, as it could become a regulatory barrier for them.

Strategically, some manufacturers would choose to register their devices first in the United States and then explore China or other markets. The United States regulation is also focused on the sponsor criteria and “Current Good Manufacturing Practice” alongside the assessment of the software algorithm itself, which makes it more organization-centric [43]. Another registration strategy could be to register half of the medical devices that are in the development stage (also called “pipeline products”) in the United States and the other half in China. After getting feedback from both authorities, they can switch them over. In the United States, applicants of new devices can go through the De Novo premarket pathway or Breakthrough Device designation to register their technology [43]. In China, there exists a “Green Channel” for software with urgent medical needs.

It is imperative for international manufacturers and regulatory authorities to engage in collaborative endeavors aimed at delineating optimal regulatory pathways for each AI and ML product. Establishing a conducive environment where stakeholders can engage in reciprocal learning is of paramount importance. Enhanced comprehension of regional regulatory variations serves as a catalyst for fostering an environment conducive to mutual learning and collaboration. Bolstering global regulatory awareness in the health care technology sphere has the potential to catalyze new opportunities, ultimately yielding enhanced benefits for patients in the long term.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Medical Device Software Information in NMPA Translated from Medical Device Classification Catalog. NMPA: National Medical Products Administration.

[[PDF File \(Adobe PDF File\), 156 KB - ai_v3i1e46871_app1.pdf](#)]

References

1. Wang X, Peng Y, Lu L, Lu Z, Bagheri M, Summers R. ChestX-Ray8: hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. 2017 Presented at: IEEE Conference on Computer Vision and Pattern Recognition (CVPR); June 27, 2024; US p. 3462-3471. [doi: [10.1109/cvpr.2017.369](https://doi.org/10.1109/cvpr.2017.369)]
2. Patel N, Michelini V, Snell J, Balu S, Hoyle AP, Parker JS, et al. Enhancing next-generation sequencing-guided cancer care through cognitive computing. *Oncologist* 2018;23(2):179-185 [FREE Full text] [doi: [10.1634/theoncologist.2017-0170](https://doi.org/10.1634/theoncologist.2017-0170)] [Medline: [29158372](https://pubmed.ncbi.nlm.nih.gov/29158372/)]
3. Ting DSW, Pasquale LR, Peng L, Campbell JP, Lee AY, Raman R, et al. Artificial intelligence and deep learning in ophthalmology. *Br J Ophthalmol* 2019;103(2):167-175 [FREE Full text] [doi: [10.1136/bjophthalmol-2018-313173](https://doi.org/10.1136/bjophthalmol-2018-313173)] [Medline: [30361278](https://pubmed.ncbi.nlm.nih.gov/30361278/)]
4. Szolovits P. *Artificial Intelligence and Medicine*. Routledge 2019:19. [doi: [10.4324/9780429052071](https://doi.org/10.4324/9780429052071)]
5. Forlenza GP. Use of artificial intelligence to improve diabetes outcomes in patients using multiple daily injections therapy. *Diabetes Technol Ther* 2019;21(S2):S24-S28 [FREE Full text] [doi: [10.1089/dia.2019.0077](https://doi.org/10.1089/dia.2019.0077)] [Medline: [31169433](https://pubmed.ncbi.nlm.nih.gov/31169433/)]
6. Tzelios C, Nathavitharana RR. Can AI technologies close the diagnostic gap in tuberculosis? *Lancet Digit Health* 2021;3(9):e535-e536 [FREE Full text] [doi: [10.1016/S2589-7500\(21\)00142-4](https://doi.org/10.1016/S2589-7500(21)00142-4)] [Medline: [34446263](https://pubmed.ncbi.nlm.nih.gov/34446263/)]
7. Notice of the state council on the new generation artificial intelligence plan. State Council of the People's Republic of China. URL: http://www.gov.cn/zhengce/content/2017-07/20/content_5211996.htm [accessed 2024-06-27]
8. Morrison WM. *China's Economic Rise: History, Trends, Challenges, and Implications for the United States*. Washington, DC: Congressional Research Service; 2013.
9. China Population in 2022. Worldometer. URL: <https://www.worldometers.info/world-population/china-population/> [accessed 2024-06-27]
10. Chinese medical device industry: how to thrive in an increasingly competitive market? Deloitte. 2021. URL: <https://www2.deloitte.com/cn/en/pages/life-sciences-and-healthcare/articles/chinese-medical-device-industry-whitepaper.html> [accessed 2024-06-27]
11. Ceross A, Bergmann J. Evaluating the presence of software-as-a-medical-device in the Australian therapeutic goods register. *Prosthesis* 2021;3(3):221-228. [doi: [10.3390/prosthesis3030022](https://doi.org/10.3390/prosthesis3030022)]
12. Announcement of the NMPA on issuing the guiding principles for technical review of medical device software registration. NMPA. URL: <https://www.nmpa.gov.cn/ylqx/ylqxgggtg/ylqxqtgg/20150805120001562.html?type=pc&m=> [accessed 2024-06-27]
13. Key points of deep learning decision-assisting medical device software review. Artificial Intelligence Medical Device Innovation and Cooperation Platform. URL: <http://www.aimd.org.cn/> [accessed 2024-06-27]
14. Announcement of NMPA on issuing the guiding principles for the classification and definition of artificial intelligence medical software products (no. 47 of 2021). NMPA. URL: <https://www.nmpa.gov.cn/xxgk/ggtg/qtggtg/20210708111147171.html?type=pc&m=> [accessed 2024-06-27]
15. Announcement of the center for device evaluation of the state food and drug administration on the release of the guidelines for the registration and review of medical device software (2022 revision) (2022 no. 9). CCFDIE. URL: <https://www.ccfdie.org/en/index.htm> [accessed 2024-06-27]
16. Notice of the center for device evaluation of the NMPA on issuing the guiding principles for the registration review of artificial intelligence medical devices (no. 8, 2022). CIRS. URL: <https://www.cirs-group.com/cn/md/gjyjjqsxgyfbrgznylqxczsczdyzdtg-2022nd8h> [accessed 2024-06-27]
17. Software as a medical device (SaMD)). FDA. URL: <https://www.fda.gov/medical-devices/digital-health-center-excellence/software-medical-device-samd> [accessed 2024-06-27]
18. Announcement of the general administration on matters concerning the implementation of the catalogue of medical devices. NMPA. URL: <https://www.nmpa.gov.cn/directory/web/nmpa/ylqx/ylqxgggtg/ylqxqtgg/20170904143301827.html> [accessed 2024-06-27]
19. Soliman E, Mogefors D, Bergmann JHM. Problem-driven innovation models for emerging technologies. *Health Technol* 2020;10(5):1195-1206. [doi: [10.1007/s12553-020-00450-5](https://doi.org/10.1007/s12553-020-00450-5)]
20. Maci J, Marešová P. Critical factors and economic methods for regulatory impact assessment in the medical device industry. *Risk Manag Healthc Policy* 2022;15:71-91 [FREE Full text] [doi: [10.2147/RMHP.S346928](https://doi.org/10.2147/RMHP.S346928)] [Medline: [35082542](https://pubmed.ncbi.nlm.nih.gov/35082542/)]

21. Arnould A, Hendricusdottir R, Bergmann J. The complexity of medical device regulations has increased, as assessed through data-driven techniques. *Prosthesis* 2021;3(4):314-330. [doi: [10.3390/prosthesis3040029](https://doi.org/10.3390/prosthesis3040029)]
22. Regulations on the supervision and administration of medical devices. NPMA. URL: http://www.gov.cn/zhengce/content/2021-03/18/content_5593739.html [accessed 2024-06-27]
23. YY/t 0664-2008 medical device software life cycle process. NMPA. 2020. URL: <https://www.codeofchina.com/standard/YYT0664-2008.html> [accessed 2024-06-27]
24. Shankui R, Xiao J, Jian F, Chunqing Z, Xinhua Y. Research on classification management of computer aided diagnosis software products. *Chin J Med Devices* 2019;5.
25. IMDRF. Software as a Medical Device: Possible Framework for Risk Categorization and Corresponding Considerations. Published online. 2014 Sep 18. URL: <https://www.imdrf.org/documents/software-medical-device-possible-framework-risk-categorization-and-corresponding-considerations> [accessed 2014-09-18]
26. National medical products administration database. CDE. URL: <http://english.nmpa.gov.cn/database.html> [accessed 2024-06-27]
27. Pashkov V, Harkusha A. Certain aspects on medical devices software law regulation. *Wiad Lek* 2016;69(6):765-767. [Medline: [28214812](https://pubmed.ncbi.nlm.nih.gov/28214812/)]
28. Pashkov VM, Harkusha AO, Harkusha YO. Artificial intelligence in medical practice: regulative issues and perspectives. *Wiad Lek* 2020;73(12):2722-2727. [doi: [10.36740/wlek202012204](https://doi.org/10.36740/wlek202012204)]
29. Proposed regulatory framework for modifications to artificial intelligence/machine learning (AI/ML)-based software as a medical device(SaMD)). FDA. URL: <https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-software-medical-device> [accessed 2024-06-27]
30. Digital health innovation action plan. FDA. URL: <https://www.fda.gov/media/106331/download/> [accessed 2024-06-27]
31. Dalton-Brown S. The ethics of medical AI and the physician-patient relationship. *Camb Q Healthc Ethics* 2020;29(1):115-121. [doi: [10.1017/S0963180119000847](https://doi.org/10.1017/S0963180119000847)] [Medline: [31858938](https://pubmed.ncbi.nlm.nih.gov/31858938/)]
32. Artificial intelligence ethical risk analysis report. Institute ETS. URL: <https://www.dx2025.com/archives/14856.html> [accessed 2024-06-27]
33. A call for greater transparency, reproducibility in use of artificial intelligence in medicine. Harvard THC. URL: <https://tinyurl.com/2s3u7ud9> [accessed 2024-06-27]
34. Cruz M, Kurapati S, Turkyilmaz-van DVY. Software reproducibility: how to put it into practice? *OSFPREPRINTS* 2018:1-8. [doi: [10.31219/osf.io/z48cm](https://doi.org/10.31219/osf.io/z48cm)]
35. Haibe-Kains B, Adam GA, Hosny A, Khodakarami F, Massive Analysis Quality Control (MAQC) Society Board of Directors, Waldron L, et al. Transparency and reproducibility in artificial intelligence. *Nature* 2020;586(7829):E14-E16 [FREE Full text] [doi: [10.1038/s41586-020-2766-y](https://doi.org/10.1038/s41586-020-2766-y)] [Medline: [33057217](https://pubmed.ncbi.nlm.nih.gov/33057217/)]
36. Key points of deep learning decision-assisting medical device software review. Artificial Intelligence Medical Device Innovation and Cooperation Platform. URL: <http://aimd.org.cn/newsinfo/1339997.html?templateId=506998> [accessed 2024-06-27]
37. Cybersecurity. FDA. URL: <https://www.fda.gov/medical-devices/digital-health-center-excellence/cybersecurity> [accessed 2024-06-27]
38. McKeon J. Cyberattacks will be the top health tech hazard this year, ECRI says. *HEALTH IT SECURITY*. URL: <https://healthitsecurity.com/news/cyberattacks-will-be-the-top-health-tech-hazard-this-year-ecri-says> [accessed 2024-06-27]
39. Group MDCW. Principles and practices for medical device cybersecurity. In: IMDRF. 2020 Presented at: International Medical Device Regulators Forum; 2024 June 27; US.
40. Neishen A. Principles of medical device cybersecurity technical evaluation. IMDRF. URL: <https://www.secrss.com/articles/40158> [accessed 2024-06-27]
41. Yue M, Wenwen Z, Shuo P, Yiwu H, Bin L, Zhong L. IMDRF interpretation of personalized medical device terms. *China Pharm Affairs* 2019;33(1):41-44. [doi: [10.1201/b14081-31](https://doi.org/10.1201/b14081-31)]
42. NMPA launched the china drug regulatory science action plan. NMPA. URL: http://www.gov.cn/xinwen/2019-05/02/content_5388253.htm [accessed 2024-06-27]
43. He J, Baxter SL, Xu J, Xu J, Zhou X, Zhang K. The practical implementation of artificial intelligence technologies in medicine. *Nat Med* 2019;25(1):30-36 [FREE Full text] [doi: [10.1038/s41591-018-0307-0](https://doi.org/10.1038/s41591-018-0307-0)] [Medline: [30617336](https://pubmed.ncbi.nlm.nih.gov/30617336/)]

Abbreviations

- AI:** artificial intelligence
- CT:** computed tomography
- EU:** European Union
- FDA:** Food and Drug Administration
- GDPR:** general data protection regulation
- IMDRF:** International Medical Device Regulators Forum
- MDSW:** medical device software

ML: machine learning

NMPA: National Medical Products Administration

SaMD: software as medical device

SiMD: software in a medical device

Edited by K El Emam, B Malin; submitted 01.03.23; peer-reviewed by Z Yu, T Zhang; comments to author 19.07.23; revised version received 20.10.23; accepted 16.06.24; published 29.07.24.

Please cite as:

Han Y, Ceross A, Bergmann J

Regulatory Frameworks for AI-Enabled Medical Device Software in China: Comparative Analysis and Review of Implications for Global Manufacturer

JMIR AI 2024;3:e46871

URL: <https://ai.jmir.org/2024/1/e46871>

doi: [10.2196/46871](https://doi.org/10.2196/46871)

PMID: [39073860](https://pubmed.ncbi.nlm.nih.gov/39073860/)

©Yu Han, Aaron Ceross, Jeroen Bergmann. Originally published in JMIR AI (<https://ai.jmir.org>), 29.07.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR AI, is properly cited. The complete bibliographic information, a link to the original publication on <https://www.ai.jmir.org/>, as well as this copyright and license information must be included.

Viewpoint

Toward Clinical Generative AI: Conceptual Framework

Nicola Luigi Bragazzi¹, MPH, MD, PhD; Sergio Garbarino², MD, PhD

¹Human Nutrition Unit, Department of Food and Drugs, University of Parma, Parma, Italy

²Department of Neuroscience, Rehabilitation, Ophthalmology, Genetics and Maternal/Child Sciences, University of Genoa, Genoa, Italy

Corresponding Author:

Nicola Luigi Bragazzi, MPH, MD, PhD

Human Nutrition Unit, Department of Food and Drugs

University of Parma

Via Volturno 39

Parma, 43125

Italy

Phone: 39 0521 903121

Email: nicolaluigi.bragazzi@unipr.it

Abstract

Clinical decision-making is a crucial aspect of health care, involving the balanced integration of scientific evidence, clinical judgment, ethical considerations, and patient involvement. This process is dynamic and multifaceted, relying on clinicians' knowledge, experience, and intuitive understanding to achieve optimal patient outcomes through informed, evidence-based choices. The advent of generative artificial intelligence (AI) presents a revolutionary opportunity in clinical decision-making. AI's advanced data analysis and pattern recognition capabilities can significantly enhance the diagnosis and treatment of diseases, processing vast medical data to identify patterns, tailor treatments, predict disease progression, and aid in proactive patient management. However, the incorporation of AI into clinical decision-making raises concerns regarding the reliability and accuracy of AI-generated insights. To address these concerns, 11 "verification paradigms" are proposed in this paper, with each paradigm being a unique method to verify the evidence-based nature of AI in clinical decision-making. This paper also frames the concept of "clinically explainable, fair, and responsible, clinician-, expert-, and patient-in-the-loop AI." This model focuses on ensuring AI's comprehensibility, collaborative nature, and ethical grounding, advocating for AI to serve as an augmentative tool, with its decision-making processes being transparent and understandable to clinicians and patients. The integration of AI should enhance, not replace, the clinician's judgment and should involve continuous learning and adaptation based on real-world outcomes and ethical and legal compliance. In conclusion, while generative AI holds immense promise in enhancing clinical decision-making, it is essential to ensure that it produces evidence-based, reliable, and impactful knowledge. Using the outlined paradigms and approaches can help the medical and patient communities harness AI's potential while maintaining high patient care standards.

(JMIR AI 2024;3:e55957) doi:[10.2196/55957](https://doi.org/10.2196/55957)

KEYWORDS

clinical intelligence; artificial intelligence; iterative process; abduction; benchmarking; verification paradigms

Clinical Decision-Making and Clinical Intelligence

Clinical decision-making can be defined as a fundamental aspect of health care practice, encompassing a wide set of skills, competencies, processes, and outcomes through which clinicians gather and analyze relevant patient data; differentiate among various conditions; and diagnose, treat, and manage patient care, balancing the effectiveness, risks, and benefits of each treatment; patient preferences; and other related values within broader societal and cultural contexts and guidelines or standards of care [1-3].

Clinical decision-making involves a complex interplay of research and biomedical knowledge, experience, and intuitive understanding developed through years of practice, contextual analytical reasoning, patient-centeredness, and compliance with ethical standards and legal requirements, with the goal of arriving at optimal health outcomes for patients by making informed, evidence-based, and shared choices while ensuring patient autonomy and confidentiality [4,5].

The 4 major pillars of clinical decision-making are scientific evidence, clinical judgment (in some complex cases not isolated to 1 clinician but involving a team of health care professionals, each contributing their expertise), ethical considerations, and

patient involvement, which are pivotal to the delivery of high-quality health care [6,7].

Clinical decision-making is not a static but rather a dynamic, multifaceted, iterative process based on reflective practice, which implies reviewing and auditing clinical decisions and outcomes to continuously learn and improve decision-making skills in the face of uncertainty and epistemic risks [5,8].

The Advent of Generative Artificial Intelligence and Its Role in Supporting Clinical Decision-Making

Artificial intelligence (AI) [9] and, in particular, generative AI [10] have the potential to revolutionize the field of clinical decision-making with their advanced capabilities in data analysis and pattern recognition. However, together with their rise, there is a growing necessity to ensure that the knowledge used and produced is evidence based and reliable. This necessity stems from the potential risks and biases associated with AI-generated insights that may not align with established medical knowledge or practices.

Generative AI can process vast amounts of medical data, including patient records, imaging data, laboratory test results, other diagnostic inputs, and clinical studies, as well as research papers, to identify patterns and correlations that might be missed by clinicians. By analyzing patient data, generative AI can help in tailoring treatments to individual patients, improving the efficacy of therapies and reducing side effects, predicting disease progression and potential complications, aiding clinicians in proactive patient management, and assisting in diagnosing

diseases, potentially identifying conditions earlier and more accurately than using traditional methods [11].

On the other hand, generative AI can produce “hallucinations” or even “fabrications” and “falsifications,” generating inaccurate or misleading information that does not accurately reflect the data it was trained on or reality [12,13], which is of particular concern in the medical realm.

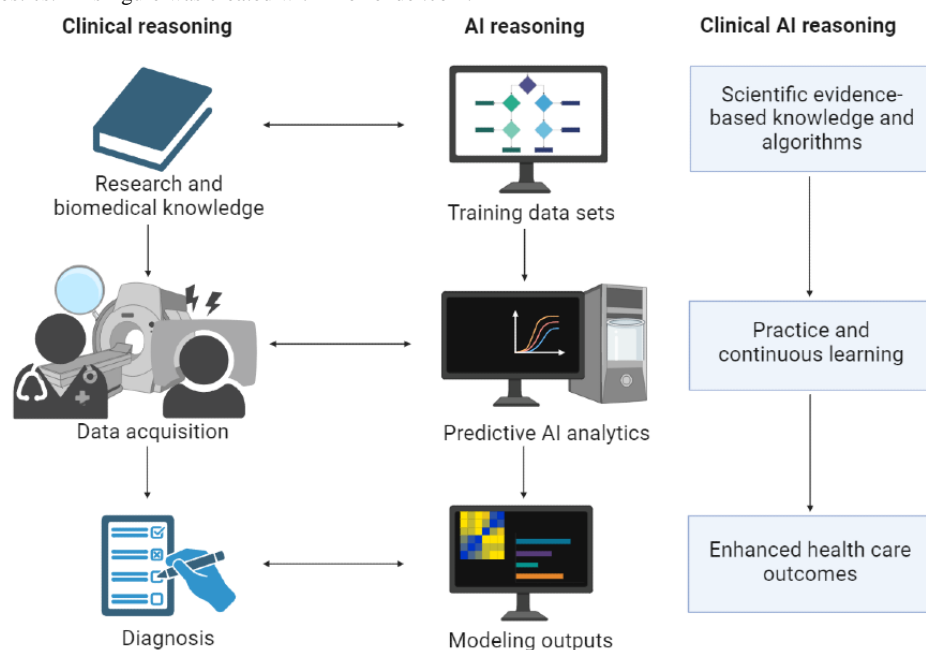
Addressing these challenges requires a multifaceted approach, including improving data set quality and diversity, refining model architectures, and incorporating mechanisms for fact checking and validation. Moreover, developing methodologies for the model to express uncertainty or request clarification when generating outputs on topics in which it has less confidence could enhance reliability. In real-world clinical applications where accuracy and truthfulness are paramount, it is crucial to implement safeguards such as human oversight, rigorous testing across diverse scenarios, and continuous monitoring and updating of AI-based models to mitigate the risks associated with these inaccuracies.

In this conceptual paper, to address these concerns, we introduce 11 “verification paradigms,” with each paradigm being a unique method to verify the evidence-based nature of AI in clinical decision-making.

Comparing Clinical Versus AI Reasoning

Interesting parallelisms between clinical decision-making and AI reasoning can be drawn (Figure 1), especially in the context of frequentist and Bayesian thinking and large language models (LLMs) such as GPT-4, which use conditional probability, revealing an interesting interplay of similarities and contrasts [5].

Figure 1. Integrating clinical expertise with artificial intelligence (AI) for enhanced health care outcomes—a schematic representation of the flow and interplay among traditional clinical reasoning, data acquisition, AI-driven predictive analytics, and the continuous learning cycle leading to improved patient care and diagnostics. This figure was created with BioRender.com.



In clinical decision-making, the reliance on scientific evidence mirrors AI’s dependence on extensive data sets for training.

Clinicians, through years of practice, develop an intuitive sense of diagnosis and treatment. Clinical reasoning often involves

abductive reasoning, which is a form of logical inference that starts with an observation or set of observations and then seeks to find the simplest and most likely explanation. In clinical practice, this means forming hypotheses based on symptoms and available data to diagnose a patient's condition. AI, particularly in fields such as machine learning and diagnostic algorithms, also frequently uses abductive reasoning—AI-based systems are, indeed, designed to analyze data, identify patterns, and make predictions or decisions based on that analysis. In many ways, this mirrors the process of abductive reasoning in which the most likely conclusion is drawn from the available information. For example, in medical diagnostics, AI-based systems might analyze patients' symptoms, medical history, and test results to suggest possible diagnoses. The aspect of human expertise underlying clinical reasoning somewhat parallels how AI-enhanced models develop a form of “intuition” from their vast training data [14,15].

When faced with complex cases, clinical decision-making often involves a collaborative approach among health care professionals, akin to the multifaceted approach of AI that integrates diverse data sources and algorithms. Ethical considerations and patient involvement are central to clinical decisions, much like how AI-based models need to be ethically aligned and user centric. Furthermore, both fields are dynamic and iterative—clinicians continually adapt their methods based on new research and patient feedback, similar to how AI-enhanced models evolve with new data and interactions.

On the AI side, traditional models often align with frequentist statistics, where the frequency of past events informs future predictions, somewhat like clinicians using epidemiological data. Modern AI, particularly in machine learning, uses Bayesian methods, updating the likelihood of outcomes with new data, reflecting how clinicians revise their hypotheses about diagnoses or treatments as new patient information comes to light. LLMs such as GPT-4 can predict outcomes based on conditional probability, which can be compared to clinicians using symptoms to predict diagnoses [16].

AI's proficiency in pattern recognition and predictive analysis also finds a parallel in clinical practice, where patterns in patient symptoms and test results are crucial for effective decision-making. However, despite these parallels, significant differences remain, with AI lacking the empathetic and deeply intuitive component inherent in human decision-making and clinicians interpreting data within a broader human context, an ability that AI has yet to fully replicate.

In essence, while there are notable similarities in the use of statistical methods and data analysis between clinical decision-making and AI reasoning, the human aspects of intuition, empathy, and ethical considerations underscore the unique characteristics of each field. The future of health care may lie in the harmonious integration of these 2 domains, leveraging the strengths of each to enhance medical care and improve patient outcomes (Figure 1).

Toward Clinical LLMs: Necessity of Verifying Evidence-Based Knowledge

However, the integration of generative AI into clinical decision-making necessitates a rigorous verification process to ensure the reliability and accuracy of the AI-generated insights. This verification is crucial because, as previously mentioned, AI-based models can sometimes generate conclusions based on flawed or biased data, leading to inaccurate or even harmful recommendations. It is essential that AI-generated advice aligns with current medical standards and best practices in addition to adhering to ethical standards, respecting patient autonomy, and ensuring equitable treatment [17,18].

Clinically oriented LLMs [19-25] such as ClinicalBERT, BlueBERT, CAML, DRG-LLaMA, GatorTronGPT, or PaLM have shown impressive capabilities, yet their application in clinical settings faces stringent requirements. Traditional methods of assessing these models' clinical knowledge often depend on automated evaluations using narrow benchmarks. To overcome these shortcomings, Singhal et al [25] recently introduced MultiMedQA, a comprehensive benchmark that merges 6 medical question-answering data sets covering a range of areas from professional medicine to consumer queries and includes HealthSearchQA, a new data set of medically related web-based search questions. This novel approach includes a human evaluation framework that examines model answers across various dimensions, namely, accuracy, understanding, reasoning, potential harm, and bias. The authors tested both PaLM and its instruction-tuned version, Flan-PaLM, on MultiMedQA. Flan-PaLM, using diverse prompting techniques, set a new standard in accuracy across all MultiMedQA multiple-choice data sets, including MedQA, MedMCQA, PubMedQA, and MMLU clinical topics, achieving a remarkable 67.6% accuracy in MedQA (US Medical Licensing Examination-style questions), which is >17% higher than the previous best. However, human assessments uncovered significant shortcomings. To address these, the authors introduced “instruction prompt tuning,” an efficient method for adapting LLMs to new domains with just a few examples. The resultant model, Med-PaLM, shows promise, yet it still does not match clinician performance even though the authors could observe that model scale and instruction prompt tuning significantly enhance comprehension, knowledge recall, and reasoning.

A further risk is that LLMs might reinforce existing biases and provide inaccurate medical diagnoses, potentially leading to detrimental effects on health care. Zack et al [26] aimed to evaluate whether GPT-4 harbors biases that could influence its application in health care settings. Using the Azure OpenAI interface, the authors scrutinized GPT-4 for racial and gender biases and assessed the impact of such biases on four clinical applications of LLMs—(1) medical education, (2) diagnostic reasoning, (3) development and implementation of clinical plans, and (4) subjective patient evaluations—involving experiments using prompts mimicking typical GPT-4 use in clinical and medical educational settings and drawing from *New England Journal of Medicine* Healer clinical vignettes and research on

implicit bias in health care. The study compared GPT-4's estimates of demographic distributions of medical conditions against actual US prevalence data. For differential diagnosis and treatment planning, the research analyzed variations across demographic groups using standard statistical methods to identify significant differences. The study revealed that GPT-4 inadequately represents demographic diversity in medical conditions, often resorting to stereotypical demographic portrayals in clinical vignettes. The differential diagnoses generated by GPT-4 for standardized clinical vignettes tended to reflect biases associated with race, ethnicity, and gender. Furthermore, the model's assessments and plans demonstrated a notable correlation between demographic characteristics and recommendations for costlier procedures, as well as varied perceptions of patients.

All this, taken together, suggests the potential role of LLMs in medicine, but human evaluations also highlight the current models' limitations, underscoring the importance of comprehensive evaluation frameworks and continued

methodological advancements to develop safe, effective LLMs for clinical use.

Implementing “Verification Paradigms”: A Comprehensive Evaluation Framework

Overview

Several “simulation and scenario testing” or “verification” paradigms can be particularly effective in verifying the evidence-based nature of generative AI in clinical decision-making. The 11 paradigms proposed in this paper were devised following thorough familiarization with existing literature and extensive consultation with experts in the field to ensure that the methodologies were not only grounded in the latest academic research and theoretical frameworks but also shaped by practical insights and recommendations from medical professionals and AI technology specialists ([Textbox 1](#) and [Table 1](#)).

Textbox 1. Overview of the verification paradigms.

Verification paradigms and brief description

- Quiz, vignette and knowledge survey: uses clinical scenarios to test artificial intelligence (AI)'s medical knowledge and reasoning.
- Historical data comparison: compares AI recommendations with known clinical outcomes to gauge accuracy.
- Expert consensus: evaluates AI-generated diagnoses or treatment plans against expert medical opinion.
- Cross-discipline validation: verifies AI insights with professionals from various medical disciplines for comprehensive evaluation.
- Rare or complex simulation and scenario testing: assesses AI's ability to handle rare and complex medical cases through simulated scenarios.
- False myth: tests AI's capability to identify and reject medical myths or outdated concepts.
- Challenging (or controversial) question: presents AI with complex medical questions to evaluate its nuanced understanding and reasoning.
- Real-time monitoring: monitors AI recommendations in clinical settings to observe real-world efficacy and safety.
- Algorithm transparency and audit: focuses on the transparency of AI's decision-making process and its ability to be audited.
- Feedback loop: involves continuous AI improvement based on feedback from practical applications and outcomes.
- Ethical and legal review: regularly reviews AI recommendations to ensure that they adhere to ethical guidelines and legal standards.

Table 1. Verification paradigms with their strengths and weaknesses.

Verification paradigm	Strengths	Weaknesses
Quiz, vignette, and knowledge survey	<ul style="list-style-type: none"> Comprehensive evaluation Real-world relevance Assessment of contextual understanding and probabilistic reasoning 	<ul style="list-style-type: none"> Complex to design Resource intensive Potential bias in test creation
Historical data comparison	<ul style="list-style-type: none"> Real-world applicability Evidence-based evaluation Objective benchmarking 	<ul style="list-style-type: none"> Dependent on data quality Historical bias May not capture AI's^a potential for novel insights
Expert consensus	<ul style="list-style-type: none"> Leverages human expertise Valuable in complex cases Incorporates ethical judgment 	<ul style="list-style-type: none"> Subjective Time-consuming Potential for expert bias
Cross-discipline validation	<ul style="list-style-type: none"> Comprehensive evaluation from multiple perspectives Mitigates the risk of siloed decision-making 	<ul style="list-style-type: none"> Coordination challenges Requires broad expert availability
Rare or complex simulation and scenario testing	<ul style="list-style-type: none"> Reveals AI's capabilities in handling diversity Can identify areas for innovation 	<ul style="list-style-type: none"> Potentially limited by available data Resource intensive
False myth	<ul style="list-style-type: none"> Tests AI's current knowledge base Assesses ability to discern evidence-based information 	<ul style="list-style-type: none"> Requires careful selection of myths Risk of reinforcing incorrect information
Challenging (or controversial) question	<ul style="list-style-type: none"> Evaluates AI's handling of ambiguity and complexity Assesses balance of different viewpoints 	<ul style="list-style-type: none"> Subjective evaluation criteria Depends on quality of input questions
Real-time monitoring	<ul style="list-style-type: none"> Direct insight into practical impact Simulates real-world testing 	<ul style="list-style-type: none"> Requires controlled clinical environment Ethical concerns with experimental use
Algorithm transparency and audit	<ul style="list-style-type: none"> Enhances trust and understanding Facilitates regulatory compliance 	<ul style="list-style-type: none"> Complexity for end users Risk of exposing proprietary information
Feedback loop	<ul style="list-style-type: none"> Ensures continuous improvement Adapts to changing medical knowledge 	<ul style="list-style-type: none"> Requires ongoing effort and resources Dependence on quality of feedback
Ethical and legal review	<ul style="list-style-type: none"> Safeguards patient rights Ensures adherence to ethical guidelines 	<ul style="list-style-type: none"> Time-consuming Needs multidisciplinary expertise

^aAI: artificial intelligence.

The Quiz, Vignette, and Knowledge Survey Paradigm

This approach involves assessing the AI's proficiency in various domains, such as medical knowledge and diagnostic reasoning, and its understanding of therapeutic interventions by using quizzes, vignettes, and validated knowledge surveys designed to mimic real-world clinical scenarios [27]. This would require the AI to have not only a vast knowledge base of medical information but also, and especially, the ability to apply this knowledge contextually, thus demonstrating an understanding of the nuances of patient presentations and how they correlate with various medical conditions and treatments. In addition, this format could incorporate elements of both frequentist and Bayesian thinking, reflecting the probabilistic nature of clinical reasoning—in other words, as previously mentioned, the AI would have to weigh the likelihood of different diagnoses based on the presented symptoms and history, similar to how clinicians use Bayesian reasoning to update their probability assessments as new information becomes available.

This approach has a number of strengths, including comprehensive evaluation, real-world relevance, contextual understanding, probabilistic reasoning assessment, and adaptability to new information. On the other hand, it suffers from some weaknesses, such as design complexity and resource intensiveness, potential bias in test creation, and lack of interdisciplinary evaluation.

Currently, this approach is the most leveraged. An extensive body of literature has found that LLMs such as ChatGPT can successfully pass medical examinations [28] although with varying degrees of heterogeneity and variability [29], exhibiting strong abilities in explanation, reasoning, memory, and accuracy. On the other hand, LLMs struggle with image-based questions [30] and, in some circumstances, lack insight and critical thinking skills [31].

Some of the studies that have exploited quizzes, vignettes, and validated knowledge surveys [32,33] have quantified the fluency and accuracy of AI-based tools using validated and reliable instruments such as the "Artificial Intelligence Performance

Instrument” [32]. This tool includes 9 items related to medical and surgical history, namely, symptoms, physical examination, diagnosis, additional examinations, management plan, and treatments. The Artificial Intelligence Performance Instrument score ranges from 0 (“inadequate clinical case management by the AI”) to 20 (“excellent clinical case management by the AI”). This score can be further subdivided into 4 subscores: patient feature, diagnosis, additional examination, and treatment score.

The Historical Data Comparison Paradigm

This approach involves comparing AI-generated recommendations with outcomes from historical data—by analyzing cases in which the clinical outcomes are well known, one can assess how well the AI’s suggestions would have aligned with actual scenarios. This would help in the comprehension of the AI’s accuracy in real-world health care settings, providing insights into its potential benefits and limitations. This is a crucial step in understanding AI’s performance and guiding its integration into clinical practice, ensuring that AI-supported decisions are in line with evidence-based medical standards and, ultimately, enhance patient care outcomes.

Strengths of this approach include real-world applicability, evidence-based evaluation, and objective benchmarking by offering a clear, objective, data-driven, and evidence-based way to benchmark AI performance against known outcomes, facilitating a straightforward and comprehensive assessment of its accuracy. Furthermore, this method enables the identification of potential gaps and improvement areas—through direct comparison with historical outcomes, specific areas in which AI recommendations may fall short can be identified, guiding further refinements. Demonstrating AI’s ability to match or surpass historical outcomes can build trust among clinicians and patients regarding AI’s utility in health care. However, this method has some weaknesses, too, including dependence on data quality in that the approach is heavily reliant on the availability and quality of historical data, with poor data quality skewing results and misleading about AI’s true performance. In addition, historical data may contain biases (eg, diagnostic, treatment, or outcome biases), which can inadvertently be reinforced by AI, affecting the fairness and accuracy of its recommendations. This shortcoming is known as “historical bias,” which arises when the data or *corpora* used to train AI-based tools no longer accurately reflect the current reality. The potential lack of novel insights is another limitation as this method benchmarks against known outcomes and may not fully capture AI’s potential to provide novel insights or diagnose conditions that were previously undetected or misdiagnosed. Furthermore, this paradigm evaluates AI against past standards of care, which may not account for advancements in medical knowledge or changes in clinical guidelines over time (“static evaluation”), and its performance on complex, multifactorial cases might not be accurately assessed if historical data are limited or if such cases were managed differently due to evolving standards of care.

Currently, to the best of our knowledge, no published studies have leveraged this approach in the biomedical arena.

The Expert Consensus Paradigm

In this paradigm, AI-generated diagnoses or treatment plans are evaluated by a panel of medical experts, with the consensus among these experts on the validity of the AI’s recommendations serving as a measure of their reliability. This paradigm is particularly useful in assessing the AI’s performance in complex cases in which human expertise is invaluable, ranging from the psychiatric field in dealing with issues such as suicide risk assessment [34] to occupational medicine [35]; oncology, with the management of malignancies [36]; and complex surgical procedures such as bariatric surgery [37].

Strengths include high-quality validation of AI’s performance, ensuring that AI-generated recommendations are thoroughly vetted by experts, and bringing a high level of scrutiny and quality control that is particularly important in complex medical fields. Incorporation of human expertise and adaptability to complex cases are other strengths by relying on medical experts to evaluate AI advice and integrating nuanced human judgment and clinical experience that AI might lack or in those instances for which AI algorithms might not have sufficient training data or might lack the capability to understand context deeply. Furthermore, expert feedback provides continuous learning opportunities, offering a platform for AI-based systems to be continuously updated and improved, enhancing their accuracy and reliability over time. This leads to heightened acceptance of AI tools as having a consensus from medical experts can increase trust among health care providers and patients in AI-generated diagnoses or treatment plans.

On the other hand, expert feedback is time and resource intensive—gathering a panel of experts and reaching a consensus can be time-consuming and expensive, which may not be feasible for every clinical decision or in settings with limited resources. In addition, despite being experts, humans are subject to biases that might affect their judgment, potentially leading to the validation of inaccurate AI recommendations. Scalability issues represent a further shortcoming—the approach may not scale well to everyday clinical practice, where quick decision-making is often required and the luxury of convening an expert panel for each AI recommendation is not practical. Furthermore, variability in expert opinion could lead to inconsistent validation of AI-generated recommendations and uncertainty in their reliability. Finally, there is a risk that this paradigm could discourage direct validation of AI algorithms through objective measures or independent verification, potentially overlooking errors or biases in the AI-based systems themselves.

The Cross-Discipline Validation Paradigm

This paradigm is rooted in the understanding that health care delivery increasingly relies on the expertise and coordination of diverse professionals to address complex health issues effectively. This approach recognizes that no single professional has all the knowledge and skills necessary to provide comprehensive care, especially in cases that involve multifaceted medical, psychological, social, and ethical considerations. As clinical decision-making is seen as a multidisciplinary teamwork process, this verification paradigm involves cross-verifying AI-generated insights with experts from various medical

disciplines. For example, a diagnosis made by an AI based on radiology images could be evaluated by experts in radiology, oncology, and pathology. This multidisciplinary approach ensures comprehensive evaluation and mitigates the risk of siloed decision-making, which is known to result in incomplete information, lack of coordination, and duplication of efforts, leading to inefficient care, higher costs, increased risk of medical errors, and decreased patient satisfaction, ultimately impacting the quality of patient care and health outcomes.

Currently, little is known about the multidisciplinary nature of LLMs. Li et al [38] evaluated the proficiency of AI-based tools in addressing interdisciplinary queries in cardio-oncology, leveraging a questionnaire consisting of 25 questions compiled based on the 2022 European Society of Cardiology guideline on cardio-oncology. ChatGPT-4 showed the highest percentage of good responses at 68%, followed by Bard, Claude 2, and ChatGPT-3.5 at 52% and LLaMA 2 at 48%. A specific area of concern was in treatment and prevention, where all LLMs scored poorly or borderline, particularly when their advice deviated from current guidelines, such as the recommendation to interrupt cancer treatment for patients with acute coronary syndrome. Other studies have assessed LLMs as support tools for multidisciplinary tumor boards in the planning of therapeutic programs for patients with cancer [39,40].

The Rare or Complex Simulation and Scenario Testing Paradigm

In this method, the AI-based tool is tested against a variety of simulated clinical scenarios, including rare and complex cases such as frail patients with multiple comorbidities, unusual presentations of diseases, or cases in which symptoms are ambiguous or misleading. This comprehensive testing can identify areas for innovation and reveal the strengths and limitations of the AI-based tool in diverse clinical situations, such as AI's capabilities in handling diversity. Conversely, this paradigm can be resource intensive and potentially limited by available data.

A recent study [41] explored ChatGPT's potential contributions to the diagnosis and management of rare and complex diseases, such as idiopathic pulmonary arterial hypertension, Klippel-Trenaunay syndrome, early-onset Parkinson disease, and Rett syndrome. LLMs can detect the disease early through AI-driven analysis of patient symptoms and medical imaging data, rapidly analyze an extensive body of biomedical literature for a better understanding of the mechanisms underlying the disease, and offer access to the latest research findings and personalized treatment plans.

Another study [42] examined the efficacy of 3 popular LLMs in medical education, particularly for diagnosing rare and complex diseases, and explored the impact of prompt engineering on their performance. Experiments were conducted on 30 cases from a diagnostic case challenge collection using various prompt strategies and a majority voting approach to compare the LLMs' performance against human consensus and MedAlpaca, an LLM designed for medical tasks. The findings revealed that all tested LLMs surpassed the average human consensus and MedAlpaca's performance by margins of at least 5% and 13%, respectively. In categories of frequently

misdiagnosed cases, Google Bard equaled MedAlpaca but exceeded human consensus by 14%. GPT-4 and GPT-3.5 showed superior performance over MedAlpaca and human respondents in often moderately misdiagnosed cases, with minimum accuracy improvements of 28% and 11%, respectively. Using a majority voting strategy, particularly with GPT-4, yielded the highest overall accuracy across the diagnostic complex case collection. On the Medical Information Mart for Intensive Care III data sets, Google Bard and GPT-4 reached the highest diagnostic accuracy scores of 93% with multiple-choice prompts, whereas GPT-3.5 and MedAlpaca scored 73% and 47%, respectively.

The False Myth Paradigm

This paradigm involves deliberately introducing known medical myths or outdated concepts into the AI's training data. The AI's ability to identify and reject these myths serves as a test of its understanding of current medical knowledge and its ability to discern evidence-based information. On the other hand, this approach requires a careful selection of myths and, if used in an inappropriate way, can reinforce incorrect information.

A few studies have harnessed this approach [43,44]. These studies evaluated the accuracy of 2 AI tools, ChatGPT-4 and Google Bard, in debunking 20 sleep-related myths using a 5-point Likert scale for falseness and public health significance and compared their performance with expert opinions. ChatGPT labeled 85% of the statements as either "false" (45%) or "generally false" (40%), showing high reliability in identifying inaccuracies, especially regarding sleep myths surrounding timing, duration, and behaviors during sleep. The tool demonstrated varying success in other categories such as presleep behaviors and brain function related to sleep. On a 5-point Likert scale, ChatGPT scored an average of 3.45 (SD 0.87) in identifying the falseness of statements and 3.15 (SD 0.99) in understanding their public health significance, indicating a good level of accuracy and understanding. Similarly, Google Bard identified 19 out of 20 statements as false, which was not significantly different from ChatGPT-4's accuracy. Google Bard's average falseness rating was 4.25 (SD 0.70), with skewness of -0.42 and kurtosis of -0.83, indicating a distribution with fewer extreme values compared to that of ChatGPT-4. For public health significance, Google Bard scored an average of 2.4 (SD 0.80), with skewness and kurtosis of 0.36 and -0.07, respectively, suggesting a more normal distribution than that of ChatGPT-4. The intraclass correlation coefficient between Google Bard and sleep experts was 0.58 for falseness and 0.69 for public health significance, showing moderate agreement. Text mining analysis showed that Google Bard focused on practical advice, whereas ChatGPT-4 emphasized theoretical aspects. A readability analysis found that Google Bard's responses matched an 8th-grade reading level, making them more accessible than ChatGPT-4's, which aligned with a 12th-grade level.

The Challenging (or Controversial) Question Paradigm

In this paradigm, the AI-based tool is presented with controversial or complex medical questions that do not have straightforward answers. The way in which AI navigates these questions, balancing different viewpoints and evidence, can

reveal its depth of understanding and its ability to handle nuanced medical issues. In the realm of medicine, evidence is hierarchical, with systematic reviews and meta-analyses at the top. An analytical evaluation would consider how the AI prioritizes, evaluates, and appraises different levels of evidence and whether it can differentiate between high-quality and lower-quality studies. In addition, AI should detect and minimize biases present in medical literature and data sources. Analytically, this involves evaluating the algorithms for their ability to identify potential biases in studies (eg, publication bias and selection bias) and adjust their conclusions accordingly. Shortcomings of this paradigm include subjective evaluation criteria and dependence on the quality of input questions.

A few studies [45,46] have assessed the skills of AI-based tools in understanding or generating complex and nuanced clinical documents, such as guidelines.

The Real-Time Monitoring Paradigm

In this paradigm, the AI's recommendations are implemented in a controlled clinical environment, and patient outcomes are closely monitored, simulating a randomized controlled trial (RCT). This real-world testing provides valuable feedback on the AI's efficacy and safety in actual clinical settings.

While this paradigm can provide direct insights into practical impact and simulate real-world testing, it requires a controlled clinical environment and may be limited by ethical concerns related to the experimental use of AI.

So far, only a few RCTs have been implemented. A recent blinded RCT [47] explored the efficacy of ChatGPT alongside traditional typing and dictation methods in assisting health care providers with clinical documentation, specifically in writing a history of present illness based on standardized patient histories. A total of 11 participants, including medical students, orthopedic surgery residents, and attending surgeons, were tasked with documenting history of present illness using 1 of the 3 methods for each of the 3 standardized patient histories. The methods were assessed for speed, length, and quality of documentation. Results indicated that, while dictation was the fastest method and resulted in longer and higher-quality patient histories according to the Physician Documentation Quality Instrument score, ChatGPT ranked intermediate in terms of speed. However, ChatGPT-generated documents were more comprehensive and organized than those produced through typing or dictation. A significant drawback noted was the inclusion of erroneous information in slightly more than one-third of ChatGPT-generated documents, raising concerns about accuracy. In addition, there was a lack of consensus among reviewers regarding the quality of patient histories.

In another controlled trial [48], ChatGPT's utility in providing empathetic responses to people with multiple sclerosis was assessed. The study recruited a sample of 1133 participants (mean age 45.26, SD 11.50 years; 68.49% female), who were surveyed through a web-based form distributed via digital communication platforms. Participants, blinded to the authors of the responses, evaluated alternate responses to 4 questions on a Likert scale from 1 to 5 for overall satisfaction and used the Consultation and Relational Empathy scale for assessing

perceived empathy. Results showed that ChatGPT's responses were perceived as significantly more empathetic than those from neurologists. However, there was no significant association between ChatGPT's responses and mean satisfaction. College graduates were significantly less likely to prefer ChatGPT's responses compared to those with a high school education.

The Algorithm Transparency and Audit Paradigm

This paradigm focuses on the transparency of the AI algorithms and the ability to audit their decision-making processes. By understanding how the AI-based tool arrives at its conclusions, clinicians can better assess the validity of its recommendations, which is crucial for building trust in AI-based systems among health care professionals.

Strengths include improved decision-making and enhanced trust and confidence by demystifying how decisions are made, thus building trust among clinicians and patients, crucial for the acceptance and integration of AI in health care. Clinicians can make more informed decisions by understanding the reasoning behind AI recommendations, potentially leading to better patient outcomes. AI-based tools can also facilitate regulatory compliance—transparency is key to meeting regulatory standards for medical devices and software, including AI-based systems used in health care. AI enables continuous improvement as a transparent decision-making process allows for easier identification of errors or biases in the AI system, facilitating ongoing refinement and improvement. Furthermore, exposing the decision-making process has educational benefits for health care professionals, helping them understand complex AI methodologies and enhancing their ability to work alongside AI tools. On the other hand, this approach has some weaknesses that should be acknowledged, including complexity for end users—AI decision-making processes, especially in deep learning, can be incredibly complex and difficult for end users to understand, potentially limiting the effectiveness of transparency. Understanding and trusting the AI process might lead some clinicians to overrely on AI recommendations without applying their judgment, especially in ambiguous or complex cases. Complete transparency might expose proprietary algorithms to potential theft or misuse, challenging companies to balance transparency with protecting their intellectual property. Moreover, there is potential room for misinterpretation—there is a risk that transparency could lead to misinterpretation of how AI algorithms work, especially without a strong foundation in data science or AI methodologies among health care professionals. Finally, developing transparent AI systems that are also understandable to clinicians requires significant resources, including time and expertise, potentially slowing down innovation.

The Feedback Loop Paradigm

This approach involves the continuous updating of the AI system based on feedback from its practical applications, with clinicians providing feedback on the AI's performance, which is then used to refine and improve the AI models. This iterative, ongoing process ensures that the AI-based system properly evolves and adapts to changing medical knowledge and practices. Conversely, it also requires ongoing efforts and resources in addition to depending on the quality of the feedback.

A few studies have investigated reproducibility and repeatability [49,50]. In a study [49] involving emergency physicians, 6 unique prompts were used in conjunction with 61 patient vignettes to assess ChatGPT's ability to assign Canadian Triage and Acuity Scale scores through 10,980 simulated triages. ChatGPT returned a Canadian Triage and Acuity Scale score in 99.6% of the queries. In terms of temporal reproducibility and repeatability, the study found considerable variation in the results—21% due to repeatability (using the same prompt multiple times) and 4% due to reproducibility (using different prompts). ChatGPT's overall accuracy in triaging patients was 47.5%, with an undertriage rate of 13.7% and an overtriage rate of 38.7%. Of note, providing more detailed prompts resulted in slightly greater reproducibility but did not significantly improve accuracy.

In another study [50] assessing ChatGPT's proficiency in answering frequently asked questions about endometriosis, detailed internet searches were used to compile questions, which were then aligned with the European Society of Human Reproduction and Embryology (ESHRE) guidelines. An experienced gynecologist rated ChatGPT's responses on a scale from 1 to 4. To test repeatability, each question was asked twice, with reproducibility determined by the consistency of ChatGPT's scoring within the same category for repeated questions. Of the frequently asked questions, 91.4% (n=71) were answered completely, accurately, and sufficiently by ChatGPT. The model showed the highest accuracy in addressing symptoms and diagnosis (16/17, 94% of the questions) and the lowest accuracy in treatment-related questions (13/16, 81% of the questions). Among the 40 questions related to the ESHRE guidelines, 27 (68%) were rated as grade 1, a total of 7 (18%) were rated as grade 2, and 6 (15%) were rated as grade 3. The reproducibility rate was highest (100%) for questions in the categories of prevention, symptoms and diagnosis, and complications. However, it was lowest for questions aligned with the ESHRE guidelines, at 70%.

These contrasting findings warrant further investigation.

The Ethical and Legal Review Paradigm

The “ethical and legal review paradigm” emphasizes the importance of ensuring that AI recommendations in health care settings adhere to established ethical guidelines and legal standards, which involves regular review rounds of the AI's recommendations by an ethics committee or legal team. This is particularly important in sensitive areas such as critical care, emergency management, end-of-life care, or genetic testing, where the stakes of decisions are particularly high and the moral and legal implications are significant. This approach aims to safeguard patients' rights, maintain trust in AI-assisted health care, and ensure that the implementation of AI technologies in medicine is both ethically sound and legally compliant [51,52].

The deployment of AI-based tools such as ChatGPT in sensitive fields raises, indeed, several ethical and legal concerns. One significant issue is the potential for bias in AI algorithms, which can lead to unfair or incorrect outcomes. Moreover, the use of AI in these fields touches on privacy concerns, especially with the processing of personal data. Furthermore, issues regarding

accountability and liability for malpractices and bad outcomes associated with AI-influenced LLM medical decision-making represent an emerging topic in the arena of legal medicine and, more broadly, forensic science.

These concerns underscore the need for strict ethical guidelines and robust legal frameworks governing AI use in biomedical and clinical practices, with the final goal of leveraging AI's strengths while mitigating its limitations, ensuring that it serves as a tool for progress rather than a source of bias and error [52,53].

Integrating the “Verification Paradigms”

These various paradigms for assessing AI in health care contexts underscore the multifaceted and complex nature of integrating AI technologies such as ChatGPT into medical practices. These paradigms reflect a concerted effort to evaluate AI systems' proficiency, ethical alignment, and practical utility in clinical settings comprehensively. Each of these paradigms offers a unique perspective and method for verifying the reliability and accuracy of generative AI in clinical decision-making, and they can be used in combination to provide a robust validation framework (Tables 2 and 3 and Figure 2).

It is of paramount importance to note that all these paradigms do not necessarily have the same weight or importance; their relevance can vary depending on the context, the specific health care domain, and the goals of the AI system being assessed. Integrating and combining these paradigms can provide a comprehensive, robust evaluation framework that leverages the strengths of each approach while mitigating their individual limitations.

Contextual or clinical relevance can be used to prioritize these approaches—in clinical settings in which decision-making is complex and highly nuanced (eg, oncology or psychiatry), paradigms that emphasize expert consensus and cross-discipline validation may be more critical, whereas for emerging treatments or rare diseases, paradigms focusing on simulation and scenario testing and challenging questions can be invaluable to explore AI's capacity to contribute novel insights or support rare condition management. In contexts in which AI is being directly implemented into clinical workflows and related follow-up, real-time monitoring and feedback loop paradigms become essential to ensure patient safety and system efficacy.

Combining paradigms for comprehensive evaluation requires a “layered, sequential, strategic integrative approach,” starting with broad assessments such as the quiz, vignette, and knowledge survey paradigm to gauge general knowledge and reasoning abilities, followed by more specific tests such as historical data comparison for accuracy in real-world scenarios and expert consensus for nuanced judgment calls. The cross-discipline validation paradigm can be harnessed to assess AI's recommendations from multiple professional perspectives, ensuring a holistic evaluation of AI's clinical recommendations. Throughout all stages of evaluation, the ethical and legal review paradigm is continuously applied to ensure adherence to ethical standards and legal requirements, safeguarding patient rights and data privacy.

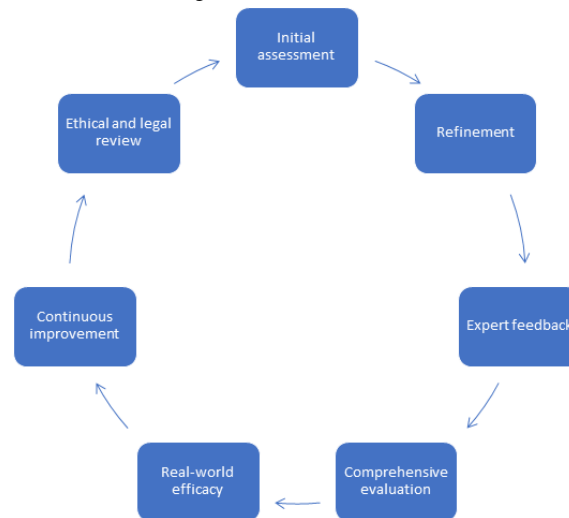
Table 2. Overview of the layered integrative approach for evaluating artificial intelligence (AI) in health care, delineating the structured, multistage framework for the comprehensive assessment and continuous improvement of AI systems.

Stage	Verification paradigm	Objective	Integration
Initial assessment	Quiz, vignette, and knowledge survey	To gauge the AI's foundational medical knowledge and its ability to apply this knowledge in simulated real-world scenarios	Forms the baseline assessment of the AI's capabilities, setting the stage for more targeted evaluations
Refinement	Historical data comparison	To refine the AI's understanding and application of medical knowledge by comparing its recommendations or diagnoses against known outcomes from historical data	Uses the insights gained from initial assessments to focus on areas requiring improvement, ensuring that the AI's recommendations are grounded in real-world evidence
Expert feedback	Expert consensus	To incorporate nuanced clinical insights and expert judgments into the AI's learning, ensuring that it aligns with current clinical practices and expert opinions	Builds on the refined knowledge base by integrating expert clinical insights, further improving the AI's decision-making processes
Comprehensive evaluation	Cross-discipline validation	To evaluate the AI's recommendations and diagnostics across various medical disciplines, ensuring a comprehensive and holistic assessment	Leverages the foundational knowledge, refined understanding, and expert insights to test the AI's capabilities in a multidisciplinary context, identifying any gaps or biases
Complexity handling	Rare or complex simulation and scenario testing	To test the AI's ability to handle complex, rare, or novel medical scenarios, ensuring that it can adapt to a wide range of clinical challenges	Uses the comprehensive evaluations as a foundation to challenge the AI with scenarios that require sophisticated reasoning, further refining its decision-making abilities
Knowledge accuracy	False myth	To ensure that the AI's current knowledge base is accurate and up-to-date, identifying and correcting any misconceptions or outdated information	Builds on the previous layers by specifically targeting and rectifying inaccuracies in the AI's knowledge, ensuring reliability
Complexity and nuance handling	Challenging (or controversial) question	To evaluate the AI's ability to navigate complex medical questions that may not have straightforward answers, assessing its reasoning in ambiguous situations	Further refines the AI's decision-making process by exposing it to nuanced clinical scenarios, enhancing its ability to provide balanced and informed recommendations
Real-world efficacy	Real-time monitoring	To monitor the AI's recommendations and diagnoses in real-world clinical settings, assessing its practical efficacy and safety	Applies all previous layers of assessment in a live clinical environment, providing direct feedback on the AI's performance and areas for improvement
Transparency and trust	Algorithm transparency and audit	To ensure that the decision-making processes of the AI are transparent and understandable, building trust among health care providers and patients	Uses insights from real-world applications and previous evaluations to demystify the AI's logic, ensuring that it is both effective and comprehensible
Continuous improvement	Feedback loop	To continuously refine and improve the AI system based on real-world data, feedback, and evolving medical knowledge	Represents the culmination of the integrative approach, in which feedback from all previous stages is used to iteratively enhance the AI system, ensuring that it remains effective, safe, and ethically compliant over time
Ethical and legal compliance	Ethical and legal review	To ensure that all AI recommendations and processes adhere to established ethical guidelines and legal standards	Runs parallel to all stages, providing a constant check on the AI's compliance with ethical norms and legal requirements, safeguarding against potential malpractices, and ensuring that patient rights are protected

Table 3. Engagement and impact of key health care stakeholders—physicians, patients, nurses, administrators, artificial intelligence (AI) developers, ethicists, and regulators—across various AI evaluation paradigms, highlighting their roles and interactions in the process of assessing and integrating AI technologies in health care.

Verification paradigm	Stakeholders						
	Physicians	Patients	Nurses	Health care administrators	AI developers	Ethicists	Regulators
Quiz, vignette, and knowledge survey	Participate in creating and testing	May be participants in scenarios	Assist in scenario design	Oversee implementation	Design relevant quizzes and surveys	Evaluate scenario ethics	Establish standards for testing
Historical data comparison	Use outcomes to validate AI	Benefit from improved outcomes	Observe AI's real-world accuracy	Use data for strategic decisions	Analyze comparison outcomes for improvement	Assess the ethical use of historical data	Monitor data use and outcomes
Expert consensus	Contribute expertise	Trust in consensus-driven AI	Support expert consensus	Involved in consensus building	Incorporate expert feedback	Participate in consensus discussions	Ensure that expert consensus meets guidelines
Cross-discipline validation	Collaborate across specialties	Benefit from holistic care approaches	Facilitate multidisciplinary care	Ensure interdisciplinary cooperation	Work with diverse health care teams	Ensure ethical cross-discipline validation	Regulate multidisciplinary validation processes
Rare or complex simulation and scenario testing	Engage in scenario creation and testing	Receive personalized care for rare conditions	Involved in patient care scenarios	Plan for innovative care solutions	Design simulations for complex conditions	Scrutinize simulations for ethical considerations	Oversee testing for safety and efficacy
False myth	Input on relevant myths	Protected from misinformation	Educate patients on myths vs facts	Promote accurate patient education	Correct and update AI knowledge	Highlight the ethical handling of myths	Regulate misinformation management
Challenging (or controversial) question	Address complex questions	Empowered by nuanced AI assistance	Assist in managing complex cases	Address policy implications	Develop algorithms for nuanced questions	Engage in ethical debates	Set standards for addressing controversial topics
Real-time monitoring	Monitor patient outcomes	Directly affected by AI recommendations	Monitor and report on patient responses	Supervise operational integration	Refine AI through real-time data	Monitor ethical implications of real-time use	Ensure patient safety in real-time monitoring
Algorithm transparency and audit	Require understanding of AI decisions	Seek transparency for trust	Advocate for clear AI explanations	Demand system transparency	Ensure algorithmic transparency	Advocate for transparent decision-making	Enforce transparency and auditability
Feedback loop	Provide clinical feedback	Benefit from ongoing improvements	Offer practical feedback	Implement system feedback	Use feedback for technical refinement	Provide ethical oversight in feedback	Facilitate regulatory feedback loops
Ethical and legal review	Ensure that AI aligns with ethical and legal standards	Protected by ethical and legal safeguards	Uphold ethical standards in AI use	Ensure compliance with regulations	Adhere to ethical and legal standards	Lead ethical and legal reviews	Conduct legal reviews and compliance checks

Figure 2. Integrating verification paradigms for artificial intelligence in health care.



This “layered, sequential, strategic integrative approach” enables continuous improvement of the entire process. An initial assessment uses paradigms such as the quiz, vignette, and knowledge survey and historical data comparison to evaluate AI’s knowledge base and practical accuracy, which are iteratively refined and optimized by applying the feedback loop paradigm using insights from real-time monitoring and expert consensus followed by algorithm transparency and audits to ensure that the system’s decisions are understandable and justifiable.

For AI-based systems targeting specific or novel medical fields, the rare or complex simulation and scenario testing should be integrated alongside challenging question paradigms to push the boundaries of AI’s capabilities and uncover areas for innovation. The feedback loop paradigm should be implemented so that AI systems are regularly updated based on new clinical evidence, shifts in expert consensus, and outcomes from real-time monitoring to ensure that AI remains aligned with current medical standards and practices through continuous evolution and adaptive learning.

This evolution is maintained transparently in terms of how feedback and new data influence AI algorithms, fostering trust among health care professionals and patients. Regular ethical and legal reviews should accompany these updates to address any emerging concerns.

Throughout the process, which is dynamic, adaptive, and iterative, a broad range of stakeholders—including patients, health care professionals, ethicists, and legal experts—should be engaged. This ensures that diverse perspectives are considered, particularly in applying paradigms such as expert consensus, ethical and legal review, and real-time monitoring. As previously mentioned, integrating these paradigms creates an ongoing process for evaluating and improving AI in health care, acknowledging the complexity of medical decision-making and the importance of maintaining ethical standards and ensuring that AI systems are not only accurate and effective but also trusted and ethical components of health care delivery.

Toward a Model of “Clinically Explainable, Fair, and Responsible Clinician-, Expert-, and Patient-in-the-Loop Artificial Intelligence”

Clinical decision-making is a cornerstone of health care, demanding a blend of knowledge, intuition, and experience. It is a dynamic process in which clinicians sift through patient data, balancing the effectiveness and risks of treatments against patient preferences and ethical standards with the goal of optimal health outcomes achieved through informed, evidence-based choices that respect patient autonomy and confidentiality [54-56].

As previously mentioned, clinical decision-making is built on 4 pillars: scientific evidence, clinical judgment, ethical considerations, and patient involvement. The integration of generative AI into this realm presents exciting possibilities and challenges—on the one hand, AI’s capacity to analyze vast amounts of medical data can enhance diagnosis, tailor treatments, and predict disease progression. However, its incorporation demands rigorous verification to align AI-generated insights with medical standards and ethical practices.

In this conceptual paper, to ensure the reliability of AI in clinical decision-making, various verification paradigms have been proposed. The quiz, vignette, and knowledge survey paradigm assesses AI’s proficiency in medical domains by using realistic scenarios to test its knowledge and contextual application incorporating frequentist and Bayesian reasoning in clinical diagnosis, whereas the historical data comparison paradigm examines AI recommendations against known clinical outcomes, assessing real-world accuracy. The expert consensus paradigm involves a panel of medical experts evaluating AI-generated diagnoses and treatment plans, whereas the cross-discipline validation paradigm cross-checks AI insights with those of professionals from different medical fields, ensuring comprehensive evaluation. In addition, the rare or complex simulation and scenario testing paradigm tests AI against a range of clinical scenarios, revealing its strengths and

limitations. The false myth paradigm tests the AI's ability to reject outdated concepts or information and content not substantiated by scientific evidence, whereas the challenging question paradigm assesses how AI handles nuanced medical issues. The real-time monitoring paradigm involves implementing AI recommendations in controlled environments to monitor patient outcomes. The algorithm transparency and audit paradigm focuses on understanding how AI reaches its conclusions, essential for clinician trust. The feedback loop paradigm ensures AI's continuous improvement based on practical application feedback. Finally, the ethical and legal review paradigm ensures that AI recommendations comply with ethical guidelines and legal requirements. Each paradigm offers a unique perspective for verifying AI in clinical decision-making, and when used in combination, they provide a comprehensive framework for ensuring the accuracy and reliability of AI, crucial for its effective integration into health care. This blend of AI and traditional clinical expertise promises a future of enhanced health care delivery, marked by precision, efficacy, and patient-centered care.

The convergence of generative AI in clinical decision-making, when rigorously verified and integrated with traditional health care practices, paves the way for a model of "clinically explainable, fair, and responsible clinician-, expert-, and patient-in-the-loop artificial intelligence." This model emphasizes not just the technical prowess of AI but also its comprehensibility, collaborative nature, and ethical grounding, ensuring that AI acts as an augmentative tool rather than an opaque, autonomous decision maker ("AI as a black box"). Clinically explainable AI demystifies the often complex and opaque decision-making processes of AI systems. In particular, the algorithm transparency and audit paradigm plays a crucial role here, ensuring that AI's reasoning is accessible and understandable to clinicians. This transparency is vital for trust and effective collaboration between human experts and AI-based systems—clinicians need to understand the rationale behind AI-generated recommendations to make informed decisions, particularly in complex or critical cases.

This understanding would also facilitate discussions and interactions with patients, who are increasingly seeking active roles in their health care decisions. By demystifying AI outputs, health care providers can offer clear, comprehensible explanations to patients, fostering trust and informed consent. Incorporating clinicians and experts in the loop is, indeed, fundamental in realizing this model—the expert consensus and cross-discipline validation paradigms highlight the importance of human expertise in evaluating and interpreting AI-generated insights, with clinicians bringing invaluable context, experience, and judgment to the table, which are crucial for nuanced decision-making. AI in this context is a tool that augments but does not replace the clinician's judgment. This collaboration ensures that AI recommendations are not only based on data and algorithms but also tempered by human insight and ethical considerations. Patient involvement is another cornerstone of this model—patient-centric care is increasingly recognized as a key component of quality health care.

The integration of AI in clinical decision-making should not diminish the patient's role but, rather, enhance it. By providing

tailored and precise medical insights, AI can empower patients with information that is specific to their condition and treatment options. This approach aligns with the growing trend toward personalized or individualized medicine, where treatments are tailored to individual patient profiles. AI can facilitate this by analyzing patient data in depth, offering insights that help with crafting personalized treatment plans. Moreover, engaging patients in the decision-making process aided by AI's insights respects their autonomy and preferences, leading to better satisfaction and adherence to treatment plans. Implementing a clinically explainable, fair, and responsible clinician-, expert-, and patient-in-the-loop AI model also necessitates continuous learning and adaptation—the feedback loop paradigm ensures that AI systems evolve based on real-world outcomes and clinician inputs. This ongoing refinement is crucial for the AI-based tool to stay relevant and effective in the ever-changing landscape of medical knowledge and practice.

Finally, the ethical and legal review paradigm ensures that AI recommendations are continually assessed for ethical and legal compliance, an aspect critical in maintaining public trust and upholding professional standards. Trust in this context extends beyond mere reliability to include ethically relevant and value-laden aspects of AI systems' design and use. This broadened understanding of trust aims to encompass concerns about fairness, transparency, privacy, and the prevention of harm, among others. While pure epistemic accounts of trust focus solely on rational and performance-based criteria, more broadly speaking, trust encompasses the full spectrum of ethical considerations necessary for truly trustworthy AI, fully integrating ethical considerations into the core of what it means for an AI system to be considered trustworthy. AI-based systems not only function effectively and reliably but also and especially operate within ethical boundaries, adhering to ethical standards and principles that respect human autonomy, prevent harm, and promote fairness and transparency [57].

In summary, the envisioned model of AI in health care is one in which AI acts as an intelligent, transparent, and adaptable assistant in the complex process of clinical decision-making, enhancing rather than replacing human expertise and keeping clinicians, experts, and patients central to the decision-making process. This approach not only leverages the strengths of AI in data processing and pattern recognition but also upholds the irreplaceable value of human judgment, experience, and ethical reasoning, all crucial for delivering high-quality patient-centered health care.

Current State of the Art and Future Directions

Currently, in a great portion of articles, the authors have limited themselves to querying the AI-based tools on a variety of topics without fully leveraging their potential. While that was understandable at the beginning of the revolution posed by LLMs, when early fascination and curiosity were prevalent, it is time to go beyond just chatting with ChatGPT and shift toward a deeper, comprehensive, and robust assessment of the capabilities of smart chatbots in real-world clinical settings. Researchers should make responsible use of AI; use standardized

reporting guidelines [58]; systematically compare different types of AI-based tools; evaluate the accuracy, repeatability, and reproducibility of the tools; and incorporate ethical and legal considerations. Validated and reliable reporting checklists are essential for ensuring that research findings and advancements are communicated clearly and consistently, facilitating comparative analyses across different AI-enhanced tools. This will help not only in identifying the most effective solutions but also in uncovering potential biases, limitations, and areas for improvement. By systematically comparing different AI-based tools and rigorously evaluating their performance, the research community can establish a benchmark for what constitutes successful integration of AI in clinical settings. A composite set of performance and outcome metrics is essential for validating the reliability of AI in clinical applications and for ensuring that tools can be confidently used across various settings without loss of performance quality. Currently, only accuracy is being investigated, with only a few studies exploring

the repeatability and reproducibility of AI-generated medical responses and recommendations.

Scholars can harness the 11 paradigms proposed in this paper to make AI-enhanced applications more clinically relevant and meaningful as well as robust and safe.

Conclusions

Generative AI holds immense promise in enhancing clinical decision-making and offering personalized, accurate, and efficient health care solutions. However, ensuring that this technology produces evidence-based, reliable, impactful knowledge is paramount. By using paradigms and approaches such as those outlined in this conceptual paper, the medical and patient communities can better leverage the potential of AI while safeguarding against misinformation and maintaining high standards of patient care.

Conflicts of Interest

None declared.

References

1. Young M, Thomas A, Lubarsky S, Ballard T, Gordon D, Gruppen LD, et al. Drawing boundaries: the difficulty in defining clinical reasoning. *Acad Med* 2018 Jul;93(7):990-995. [doi: [10.1097/ACM.0000000000002142](https://doi.org/10.1097/ACM.0000000000002142)] [Medline: [29369086](https://pubmed.ncbi.nlm.nih.gov/29369086/)]
2. Young ME, Thomas A, Lubarsky S, Gordon D, Gruppen LD, Rencic J, et al. Mapping clinical reasoning literature across the health professions: a scoping review. *BMC Med Educ* 2020 Apr 07;20(1):107 [FREE Full text] [doi: [10.1186/s12909-020-02012-9](https://doi.org/10.1186/s12909-020-02012-9)] [Medline: [32264895](https://pubmed.ncbi.nlm.nih.gov/32264895/)]
3. Benner P, Hughes RG, Sutphen M. Clinical reasoning, decisionmaking, and action: thinking critically and clinically. In: Hughes RG, editor. *Patient Safety and Quality: An Evidence-Based Handbook for Nurses*. Rockville, MD: Agency for Healthcare Research and Quality (US); 2008.
4. Andreoletti M, Berchiolla P, Boniolo G, Chiffi D. Introduction: foundations of clinical reasoning—an epistemological stance. *Topoi* 2018 Nov 30;38(2):389-394. [doi: [10.1007/s11245-018-9619-4](https://doi.org/10.1007/s11245-018-9619-4)]
5. Chiffi D. *Clinical Reasoning: Knowledge, Uncertainty, and Values in Health Care*. Cham, Switzerland: Springer International Publishing; 2020.
6. Worrall J. Evidence: philosophy of science meets medicine. *J Eval Clin Pract* 2010 Apr 30;16(2):356-362. [doi: [10.1111/j.1365-2753.2010.01400.x](https://doi.org/10.1111/j.1365-2753.2010.01400.x)] [Medline: [20367864](https://pubmed.ncbi.nlm.nih.gov/20367864/)]
7. Larson EB. How can clinicians incorporate research advances into practice? *J Gen Intern Med* 1997 Apr;12 Suppl 2(Suppl 2):S20-S24 [FREE Full text] [doi: [10.1046/j.1525-1497.12.s2.3.x](https://doi.org/10.1046/j.1525-1497.12.s2.3.x)] [Medline: [9127240](https://pubmed.ncbi.nlm.nih.gov/9127240/)]
8. Parascandola M. Epistemic risk: empirical science and the fear of being wrong. *Law Probability Risk* 2010 Jul 07;9(3-4):201-214. [doi: [10.1093/lpr/mgq005](https://doi.org/10.1093/lpr/mgq005)]
9. Müller VC. *Philosophy and Theory of Artificial Intelligence*. Berlin, Germany: Springer; 2012.
10. Schopow N, Osterhoff G, Baur D. Applications of the natural language processing tool ChatGPT in clinical practice: comparative study and augmented systematic review. *JMIR Med Inform* 2023 Nov 28;11:e48933 [FREE Full text] [doi: [10.2196/48933](https://doi.org/10.2196/48933)] [Medline: [38015610](https://pubmed.ncbi.nlm.nih.gov/38015610/)]
11. Bagde H, Dhopte A, Alam MK, Basri R. A systematic review and meta-analysis on ChatGPT and its utilization in medical and dental research. *Heliyon* 2023 Dec;9(12):e23050 [FREE Full text] [doi: [10.1016/j.heliyon.2023.e23050](https://doi.org/10.1016/j.heliyon.2023.e23050)] [Medline: [38144348](https://pubmed.ncbi.nlm.nih.gov/38144348/)]
12. Shorey S, Mattar C, Pereira TL, Choolani M. A scoping review of ChatGPT's role in healthcare education and research. *Nurse Educ Today* 2024 Apr;135:106121 [FREE Full text] [doi: [10.1016/j.nedt.2024.106121](https://doi.org/10.1016/j.nedt.2024.106121)] [Medline: [38340639](https://pubmed.ncbi.nlm.nih.gov/38340639/)]
13. Emsley R. ChatGPT: these are not hallucinations - they're fabrications and falsifications. *Schizophrenia (Heidelb)* 2023 Aug 19;9(1):52 [FREE Full text] [doi: [10.1038/s41537-023-00379-4](https://doi.org/10.1038/s41537-023-00379-4)] [Medline: [37598184](https://pubmed.ncbi.nlm.nih.gov/37598184/)]
14. Chiffi D, Zanotti R. Fear of knowledge: clinical hypotheses in diagnostic and prognostic reasoning. *J Eval Clin Pract* 2017 Oct 24;23(5):928-934. [doi: [10.1111/jep.12664](https://doi.org/10.1111/jep.12664)] [Medline: [27882636](https://pubmed.ncbi.nlm.nih.gov/27882636/)]
15. Christakis NA, Sachs GA. The role of prognosis in clinical decision making. *J Gen Intern Med* 1996 Jul;11(7):422-425. [doi: [10.1007/bf02600190](https://doi.org/10.1007/bf02600190)]

16. Savcicens G, Eliassi-Rad T, Hansen LK, Mortensen LH, Lilleholt L, Rogers A, et al. Using sequences of life-events to predict human lives. *Nat Comput Sci* 2024 Jan 18;4(1):43-56. [doi: [10.1038/s43588-023-00573-5](https://doi.org/10.1038/s43588-023-00573-5)] [Medline: [38177491](https://pubmed.ncbi.nlm.nih.gov/38177491/)]
17. Seyyed-Kalantari L, Zhang H, McDermott MB, Chen IY, Ghassemi M. Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. *Nat Med* 2021 Dec;27(12):2176-2182 [FREE Full text] [doi: [10.1038/s41591-021-01595-0](https://doi.org/10.1038/s41591-021-01595-0)] [Medline: [34893776](https://pubmed.ncbi.nlm.nih.gov/34893776/)]
18. The Lancet Digital Health. Large language models: a new chapter in digital health. *Lancet Digit Health* 2024 Jan;6(1):e1 [FREE Full text] [doi: [10.1016/S2589-7500\(23\)00254-6](https://doi.org/10.1016/S2589-7500(23)00254-6)] [Medline: [38123249](https://pubmed.ncbi.nlm.nih.gov/38123249/)]
19. Wang H, Gao C, Dantona C, Hull B, Sun J. DRG-LLaMA : tuning LLaMA model to predict diagnosis-related group for hospitalized patients. *NPJ Digit Med* 2024 Jan 22;7(1):16 [FREE Full text] [doi: [10.1038/s41746-023-00989-3](https://doi.org/10.1038/s41746-023-00989-3)] [Medline: [38253711](https://pubmed.ncbi.nlm.nih.gov/38253711/)]
20. Hwang S, Reddy S, Wainwright K, Schriver E, Cappola A, Mowery D. Using natural language processing to extract and classify symptoms among patients with thyroid dysfunction. *Stud Health Technol Inform* 2024 Jan 25;310:614-618. [doi: [10.3233/SHTI231038](https://doi.org/10.3233/SHTI231038)] [Medline: [38269882](https://pubmed.ncbi.nlm.nih.gov/38269882/)]
21. Chen F, Bokhari SM, Cato K, Gürsoy G, Rossetti S. Examining the generalizability of pretrained de-identification transformer models on narrative nursing notes. *Appl Clin Inform* 2024 Mar;15(2):357-367 [FREE Full text] [doi: [10.1055/a-2282-4340](https://doi.org/10.1055/a-2282-4340)] [Medline: [38447965](https://pubmed.ncbi.nlm.nih.gov/38447965/)]
22. Talebi S, Tong E, Li A, Yamin G, Zaharchuk G, Mofrad MR. Exploring the performance and explainability of fine-tuned BERT models for neuroradiology protocol assignment. *BMC Med Inform Decis Mak* 2024 Feb 07;24(1):40 [FREE Full text] [doi: [10.1186/s12911-024-02444-z](https://doi.org/10.1186/s12911-024-02444-z)] [Medline: [38326769](https://pubmed.ncbi.nlm.nih.gov/38326769/)]
23. Bernstein IA, Koornwinder A, Hwang HH, Wang SY. Automated recognition of visual acuity measurements in ophthalmology clinical notes using deep learning. *Ophthalmol Sci* 2024;4(2):100371 [FREE Full text] [doi: [10.1016/j.xops.2023.100371](https://doi.org/10.1016/j.xops.2023.100371)] [Medline: [37868799](https://pubmed.ncbi.nlm.nih.gov/37868799/)]
24. Peng C, Yang X, Chen A, Smith KE, PourNejatian N, Costa AB, et al. A study of generative large language model for medical research and healthcare. *NPJ Digit Med* 2023 Nov 16;6(1):210 [FREE Full text] [doi: [10.1038/s41746-023-00958-w](https://doi.org/10.1038/s41746-023-00958-w)] [Medline: [37973919](https://pubmed.ncbi.nlm.nih.gov/37973919/)]
25. Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, et al. Large language models encode clinical knowledge. *Nature* 2023 Aug;620(7972):172-180 [FREE Full text] [doi: [10.1038/s41586-023-06291-2](https://doi.org/10.1038/s41586-023-06291-2)] [Medline: [37438534](https://pubmed.ncbi.nlm.nih.gov/37438534/)]
26. Zack T, Lehman E, Suzgun M, Rodriguez JA, Celi LA, Gichoya J, et al. Assessing the potential of GPT-4 to perpetuate racial and gender biases in health care: a model evaluation study. *Lancet Digit Health* 2024 Jan;6(1):e12-e22 [FREE Full text] [doi: [10.1016/S2589-7500\(23\)00225-X](https://doi.org/10.1016/S2589-7500(23)00225-X)] [Medline: [38123252](https://pubmed.ncbi.nlm.nih.gov/38123252/)]
27. Daniel M, Rencic J, Durning SJ, Holmboe E, Santen SA, Lang V, et al. Clinical reasoning assessment methods: a scoping review and practical guidance. *Acad Med* 2019 Jun;94(6):902-912. [doi: [10.1097/ACM.0000000000002618](https://doi.org/10.1097/ACM.0000000000002618)] [Medline: [30720527](https://pubmed.ncbi.nlm.nih.gov/30720527/)]
28. Levin G, Horesh N, Brezinov Y, Meyer R. Performance of ChatGPT in medical examinations: a systematic review and a meta-analysis. *BJOG* 2024 Feb;131(3):378-380. [doi: [10.1111/1471-0528.17641](https://doi.org/10.1111/1471-0528.17641)] [Medline: [37604703](https://pubmed.ncbi.nlm.nih.gov/37604703/)]
29. Wei Q, Yao Z, Cui Y, Wei B, Jin Z, Xu X. Evaluation of ChatGPT-generated medical responses: a systematic review and meta-analysis. *J Biomed Inform* 2024 Mar;151:104620. [doi: [10.1016/j.jbi.2024.104620](https://doi.org/10.1016/j.jbi.2024.104620)] [Medline: [38462064](https://pubmed.ncbi.nlm.nih.gov/38462064/)]
30. Haver HL, Bahl M, Doo FX, Kamel PI, Parekh VS, Jeudy J, et al. Evaluation of multimodal ChatGPT (GPT-4V) in describing mammography image features. *Can Assoc Radiol J* 2024 Apr 06:8465371241247043 (forthcoming). [doi: [10.1177/08465371241247043](https://doi.org/10.1177/08465371241247043)] [Medline: [38581353](https://pubmed.ncbi.nlm.nih.gov/38581353/)]
31. Sumbal A, Sumbal R, Amir A. Can ChatGPT-3.5 pass a medical exam? A systematic review of ChatGPT's performance in academic testing. *J Med Educ Curric Dev* 2024;11:23821205241238641 [FREE Full text] [doi: [10.1177/23821205241238641](https://doi.org/10.1177/23821205241238641)] [Medline: [38487300](https://pubmed.ncbi.nlm.nih.gov/38487300/)]
32. Lechien JR, Maniaci A, Gengler I, Hans S, Chiesa-Estomba CM, Vaira LA. Validity and reliability of an instrument evaluating the performance of intelligent chatbot: the Artificial Intelligence Performance Instrument (AIPI). *Eur Arch Otorhinolaryngol* 2024 Apr;281(4):2063-2079. [doi: [10.1007/s00405-023-08219-y](https://doi.org/10.1007/s00405-023-08219-y)] [Medline: [37698703](https://pubmed.ncbi.nlm.nih.gov/37698703/)]
33. Dronkers EA, Geneid A, Al Yaghchi C, Lechien JR. Evaluating the potential of AI chatbots in treatment decision-making for acquired bilateral vocal fold paralysis in adults. *J Voice* 2024 Apr 06:S0892-1997(24)00059-6 [FREE Full text] [doi: [10.1016/j.jvoice.2024.02.020](https://doi.org/10.1016/j.jvoice.2024.02.020)] [Medline: [38584026](https://pubmed.ncbi.nlm.nih.gov/38584026/)]
34. Elyoseph Z, Levkovich I. Beyond human expertise: the promise and limitations of ChatGPT in suicide risk assessment. *Front Psychiatry* 2023 Aug 1;14:1213141 [FREE Full text] [doi: [10.3389/fpsyt.2023.1213141](https://doi.org/10.3389/fpsyt.2023.1213141)] [Medline: [37593450](https://pubmed.ncbi.nlm.nih.gov/37593450/)]
35. Padovan M, Cosci B, Petillo A, Nerli G, Porciatti F, Scarinci S, et al. ChatGPT in occupational medicine: a comparative study with human experts. *Bioengineering (Basel)* 2024 Jan 06;11(1):57 [FREE Full text] [doi: [10.3390/bioengineering11010057](https://doi.org/10.3390/bioengineering11010057)] [Medline: [38247934](https://pubmed.ncbi.nlm.nih.gov/38247934/)]
36. Peng W, Feng Y, Yao C, Zhang S, Zhuo H, Qiu T, et al. Evaluating AI in medicine: a comparative analysis of expert and ChatGPT responses to colorectal cancer questions. *Sci Rep* 2024 Feb 03;14(1):2840 [FREE Full text] [doi: [10.1038/s41598-024-52853-3](https://doi.org/10.1038/s41598-024-52853-3)] [Medline: [38310152](https://pubmed.ncbi.nlm.nih.gov/38310152/)]

37. Jazi AH, Mahjoubi M, Shahabi S, Alqahtani AR, Haddad A, Pazouki A, et al. Bariatric evaluation through AI: a survey of expert opinions versus ChatGPT-4 (BETA-SEOV). *Obes Surg* 2023 Dec;33(12):3971-3980. [doi: [10.1007/s11695-023-06903-w](https://doi.org/10.1007/s11695-023-06903-w)] [Medline: [37889368](https://pubmed.ncbi.nlm.nih.gov/37889368/)]
38. Li P, Zhang X, Zhu E, Yu S, Sheng B, Tham YC, et al. Potential multidisciplinary use of large language models for addressing queries in cardio-oncology. *J Am Heart Assoc* 2024 Mar 19;13(6):e033584 [FREE Full text] [doi: [10.1161/JAHA.123.033584](https://doi.org/10.1161/JAHA.123.033584)] [Medline: [38497458](https://pubmed.ncbi.nlm.nih.gov/38497458/)]
39. Lukac S, Dayan D, Fink V, Leinert E, Hartkopf A, Veselinovic K, et al. Evaluating ChatGPT as an adjunct for the multidisciplinary tumor board decision-making in primary breast cancer cases. *Arch Gynecol Obstet* 2023 Dec;308(6):1831-1844 [FREE Full text] [doi: [10.1007/s00404-023-07130-5](https://doi.org/10.1007/s00404-023-07130-5)] [Medline: [37458761](https://pubmed.ncbi.nlm.nih.gov/37458761/)]
40. Vela Ulloa J, King Valenzuela S, Riquoir Altamirano C, Urrejola Schmied G. Artificial intelligence-based decision-making: can ChatGPT replace a multidisciplinary tumour board? *Br J Surg* 2023 Oct 10;110(11):1543-1544. [doi: [10.1093/bjs/znad264](https://doi.org/10.1093/bjs/znad264)] [Medline: [37595064](https://pubmed.ncbi.nlm.nih.gov/37595064/)]
41. Zheng Y, Sun X, Feng B, Kang K, Yang Y, Zhao A, et al. Rare and complex diseases in focus: ChatGPT's role in improving diagnosis and treatment. *Front Artif Intell* 2024;7:1338433 [FREE Full text] [doi: [10.3389/frai.2024.1338433](https://doi.org/10.3389/frai.2024.1338433)] [Medline: [38283995](https://pubmed.ncbi.nlm.nih.gov/38283995/)]
42. Abdullahi T, Singh R, Eickhoff C. Learning to make rare and complex diagnoses with generative AI assistance: qualitative study of popular large language models. *JMIR Med Educ* 2024 Feb 13;10:e51391 [FREE Full text] [doi: [10.2196/51391](https://doi.org/10.2196/51391)] [Medline: [38349725](https://pubmed.ncbi.nlm.nih.gov/38349725/)]
43. Bragazzi NL, Garbarino S. Assessing the accuracy of generative conversational artificial intelligence in debunking sleep health myths: mixed methods comparative study with expert analysis. *JMIR Form Res* 2024 Apr 16;8:e55762 [FREE Full text] [doi: [10.2196/55762](https://doi.org/10.2196/55762)] [Medline: [38501898](https://pubmed.ncbi.nlm.nih.gov/38501898/)]
44. Garbarino S, Bragazzi NL. Evaluating the effectiveness of artificial intelligence-based tools in detecting and understanding sleep health misinformation: comparative analysis using Google Bard and OpenAI ChatGPT-4. *J Sleep Res* 2024 Apr 05:e14210. [doi: [10.1111/jsr.14210](https://doi.org/10.1111/jsr.14210)] [Medline: [38577714](https://pubmed.ncbi.nlm.nih.gov/38577714/)]
45. Saturno MP, Mejia MR, Wang A, Kwon D, Oleru O, Seyidova N, et al. Generative artificial intelligence fails to provide sufficiently accurate recommendations when compared to established breast reconstruction surgery guidelines. *J Plast Reconstr Aesthet Surg* 2023 Nov;86:248-250. [doi: [10.1016/j.bjps.2023.09.030](https://doi.org/10.1016/j.bjps.2023.09.030)] [Medline: [37793197](https://pubmed.ncbi.nlm.nih.gov/37793197/)]
46. Zaidat B, Shrestha N, Rosenberg AM, Ahmed W, Rajjoub R, Hoang T, et al. Performance of a large language model in the generation of clinical guidelines for antibiotic prophylaxis in spine surgery. *Neurospine* 2024 Mar;21(1):128-146 [FREE Full text] [doi: [10.14245/ns.2347310.655](https://doi.org/10.14245/ns.2347310.655)] [Medline: [38569639](https://pubmed.ncbi.nlm.nih.gov/38569639/)]
47. Baker HP, Dwyer E, Kalidoss S, Hynes K, Wolf J, Strelzow JA. ChatGPT's ability to assist with clinical documentation: a randomized controlled trial. *J Am Acad Orthop Surg* 2024 Feb 01;32(3):123-129. [doi: [10.5435/JAAOS-D-23-00474](https://doi.org/10.5435/JAAOS-D-23-00474)] [Medline: [37976385](https://pubmed.ncbi.nlm.nih.gov/37976385/)]
48. Maida E, Moccia M, Palladino R, Borriello G, Affinito G, Clerico M, et al. ChatGPT vs. neurologists: a cross-sectional study investigating preference, satisfaction ratings and perceived empathy in responses among people living with multiple sclerosis. *J Neurol* 2024 Apr 03 (forthcoming). [doi: [10.1007/s00415-024-12328-x](https://doi.org/10.1007/s00415-024-12328-x)] [Medline: [38568227](https://pubmed.ncbi.nlm.nih.gov/38568227/)]
49. Franc JM, Cheng L, Hart A, Hata R, Hertelendy A. Repeatability, reproducibility, and diagnostic accuracy of a commercial large language model (ChatGPT) to perform emergency department triage using the Canadian triage and acuity scale. *CJEM* 2024 Jan;26(1):40-46. [doi: [10.1007/s43678-023-00616-w](https://doi.org/10.1007/s43678-023-00616-w)] [Medline: [38206515](https://pubmed.ncbi.nlm.nih.gov/38206515/)]
50. Ozgor BY, Simavi MA. Accuracy and reproducibility of ChatGPT's free version answers about endometriosis. *Int J Gynaecol Obstet* 2024 May;165(2):691-695. [doi: [10.1002/ijgo.15309](https://doi.org/10.1002/ijgo.15309)] [Medline: [38108232](https://pubmed.ncbi.nlm.nih.gov/38108232/)]
51. Shumway DO, Hartman HJ. Medical malpractice liability in large language model artificial intelligence: legal review and policy recommendations. *J Osteopath Med* 2024 Jan 31 (forthcoming) [FREE Full text] [doi: [10.1515/jom-2023-0229](https://doi.org/10.1515/jom-2023-0229)] [Medline: [38295300](https://pubmed.ncbi.nlm.nih.gov/38295300/)]
52. Guleria A, Krishan K, Sharma V, Kanchan T. ChatGPT: forensic, legal, and ethical issues. *Med Sci Law* 2024 Apr;64(2):150-156. [doi: [10.1177/00258024231191829](https://doi.org/10.1177/00258024231191829)] [Medline: [37528607](https://pubmed.ncbi.nlm.nih.gov/37528607/)]
53. Amram B, Klempner U, Shturman S, Greenbaum D. Therapists or replicants? Ethical, legal, and social considerations for using ChatGPT in therapy. *Am J Bioeth* 2023 May;23(5):40-42. [doi: [10.1080/15265161.2023.2191022](https://doi.org/10.1080/15265161.2023.2191022)] [Medline: [37130418](https://pubmed.ncbi.nlm.nih.gov/37130418/)]
54. Hood L, Auffray C. Participatory medicine: a driving force for revolutionizing healthcare. *Genome Med* 2013;5(12):110 [FREE Full text] [doi: [10.1186/gm514](https://doi.org/10.1186/gm514)] [Medline: [24360023](https://pubmed.ncbi.nlm.nih.gov/24360023/)]
55. Gorini A, Pravettoni G. P5 medicine: a plus for a personalized approach to oncology. *Nat Rev Clin Oncol* 2011 May 31;8(7):444. [doi: [10.1038/nrclinonc.2010.227-c1](https://doi.org/10.1038/nrclinonc.2010.227-c1)] [Medline: [21629214](https://pubmed.ncbi.nlm.nih.gov/21629214/)]
56. Bragazzi NL. From P0 to P6 medicine, a model of highly participatory, narrative, interactive, and "augmented" medicine: some considerations on Salvatore Iaconesi's clinical story. *Patient Prefer Adherence* 2013;7:353-359 [FREE Full text] [doi: [10.2147/PPA.S38578](https://doi.org/10.2147/PPA.S38578)] [Medline: [23650443](https://pubmed.ncbi.nlm.nih.gov/23650443/)]
57. Zanotti G, Petrolo M, Chiffi D, Schiaffonati V. Keep trusting! A plea for the notion of trustworthy AI. *AI & Soc* 2023 Oct 12. [doi: [10.1007/s00146-023-01789-9](https://doi.org/10.1007/s00146-023-01789-9)]
58. Cacciamani GE, Collins GS, Gill IS. ChatGPT: standard reporting guidelines for responsible use. *Nature* 2023 Jun;618(7964):238. [doi: [10.1038/d41586-023-01853-w](https://doi.org/10.1038/d41586-023-01853-w)] [Medline: [37280286](https://pubmed.ncbi.nlm.nih.gov/37280286/)]

Abbreviations

AI: artificial intelligence

ESHRE: European Society of Human Reproduction and Embryology

LLM: large language model

RCT: randomized controlled trial

Edited by K El Emam, Y Zhuang; submitted 30.12.23; peer-reviewed by D Chiffi, M Andreatti, L Zhu; comments to author 13.03.24; revised version received 08.04.24; accepted 06.05.24; published 07.06.24.

Please cite as:

Bragazzi NL, Garbarino S

Toward Clinical Generative AI: Conceptual Framework

JMIR AI 2024;3:e55957

URL: <https://ai.jmir.org/2024/1/e55957>

doi: [10.2196/55957](https://doi.org/10.2196/55957)

PMID: [38875592](https://pubmed.ncbi.nlm.nih.gov/38875592/)

©Nicola Luigi Bragazzi, Sergio Garbarino. Originally published in JMIR AI (<https://ai.jmir.org>), 07.06.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR AI, is properly cited. The complete bibliographic information, a link to the original publication on <https://www.ai.jmir.org/>, as well as this copyright and license information must be included.

Viewpoint

The Utility and Implications of Ambient Scribes in Primary Care

Puneet Seth^{1*}, BSc, MD; Romina Carretas^{2*}, MPH; Frank Rudzicz^{3,4*}, PhD

¹Department of Family Medicine, McMaster University, Hamilton, ON, Canada

²School of Public Health, University of Alberta, Edmonton, AB, Canada

³Faculty of Computer Science, Dalhousie University, Halifax, NS, Canada

⁴Vector Institute for Artificial Intelligence, Toronto, ON, Canada

* all authors contributed equally

Corresponding Author:

Puneet Seth, BSc, MD

Department of Family Medicine

McMaster University

100 Main Street West

Hamilton, ON, L8P 1H6

Canada

Phone: 1 416 671 5114

Email: sethp1@mcmaster.ca

Abstract

Ambient scribe technology, utilizing large language models, represents an opportunity for addressing several current pain points in the delivery of primary care. We explore the evolution of ambient scribes and their current use in primary care. We discuss the suitability of primary care for ambient scribe integration, considering the varied nature of patient presentations and the emphasis on comprehensive care. We also propose the stages of maturation in the use of ambient scribes in primary care and their impact on care delivery. Finally, we call for focused research on safety, bias, patient impact, and privacy in ambient scribe technology, emphasizing the need for early training and education of health care providers in artificial intelligence and digital health tools.

(JMIR AI 2024;3:e57673) doi:[10.2196/57673](https://doi.org/10.2196/57673)

KEYWORDS

artificial intelligence; AI; large language model; LLM; digital scribe; ambient scribe; organizational efficiency; electronic health record; documentation burden; administrative burden

Introduction

Integrating artificial intelligence (AI) in health care has opened new horizons for improving clinical efficiency and patient care. Given the integral role that communication plays in all aspects of clinical care, particularly during patient-physician conversation, using AI to enhance communication and reduce workflow friction has immense implications. Ambient scribes are AI-powered systems that passively listen to and analyze health care provider-patient conversations, automatically generating accurate clinical documentation. Leveraging automatic speech recognition and modern forms of AI, ambient scribes stand at the forefront of the health AI revolution [1].

Large language models (LLMs), a form of AI trained on massive amounts of data that can generate text and respond to requests as if they understand them, have been a recent catalyst in the capabilities of ambient scribes. Initially, automatic speech recognition demonstrated moderate accuracy in converting

speech to text and lacked contextual understanding [2]. However, more modern neural network models such as ClinicalBERT [3], leveraging components based on transformer networks, offer more nuanced understanding and text generation nearly indistinguishable from human performance [4]. The internal mechanisms of these transformers, including self-attention components, may enable models to discern relevant parts of conversations, which is essential in complex health care dialogues [5]. Moreover, recent LLMs such as GPT-4, developed with reinforcement learning, have shown capabilities beyond traditional models, including passing scores on all steps of the USMLE (United States Medical Licensing Examination), demonstrating understanding and potential across medical contexts [4].

Challenges in Primary Care

In contemporary health care, primary care is experiencing an acute strain, arguably more so than other medical disciplines

[6]. The sector is grappling with significant challenges, most notably health care provider burnout and an escalating human resource crisis [6]. In Canada, 2023 marked an unprecedented trend with a record number of unfilled positions in primary care residency programs [7]. Concurrently, there has been an alarming increase in the number of primary care providers leaving the profession, a phenomenon partly attributable to the overwhelming administrative burdens they face [6]. Primary care, characterized by its multifaceted nature—commonly encompassing multi-issue visits, ambiguous clinical presentations, and a diverse array of visit types—demands significant administrative work from providers [8]. This, combined with the inherently unique characteristics of primary care consultations, positions this domain to benefit from the adoption of ambient scribes. By alleviating some of the administrative pressures, ambient scribes may significantly mitigate these pain points, offering hope for an overburdened primary care system.

While comprehensive data on ambient scribe use in health care is sparse, anecdotal evidence suggests a growing adoption in primary care [9]. These tools have shown potential in reducing the administrative burden, allowing clinicians to focus more on patient care. This shift is particularly evident in primary care, where the diversity and ambiguity of clinical presentations demands flexible and efficient documentation methods [10].

The Stages of Maturation of Ambient Scribe Use in Primary Care

The advancement of ambient scribe utilization within primary care can be described in a staged process based on the nature of the activities that are supported by the tool. We posit four high-level stages, shown in Table 1. The rationale behind the four stages is based on an ascending degree of complexity associated with several factors, including technical complexity in development, medicolegal barriers to adoption, and cultural factors in the practice of medicine that would impact adoption [5,11].

Table 1. Key activities associated with various stages of ambient scribe maturation in a clinical setting.

Key activity	Stage 1	Stage 2	Stage 3	Stage 4
Automation of clinical documentation	✓	✓	✓	✓
Automation of administrative actions		✓	✓	✓
Reactive clinical decision support			✓	✓
Proactive clinical decision support				✓

Stage 1 describes the most basic ambient scribe functionality, in which the tools exclusively automate clinical documentation. This may involve integration with an electronic medical record (EMR) and typically does not require information retrieval from the EMR. Stage 2 adds the ability of the ambient scribe to address administrative workflow improvements for the clinician, such as generating a letter, filling out a form, or generating tasks to be completed. Most present-day ambient scribes are likely in stages 1 or 2.

Stage 3 introduces the first clinical decision support capabilities of the ambient scribe. These would be reactive, in that they would be initiated by the clinician. For example, the clinician could consult the ambient scribe with a clinical question, such as asking about the dosing of a medication or other diagnostic possibilities. This would necessitate that the ambient scribe has access to medical knowledge and has been trained for this purpose.

Lastly, stage 4, which we imagine to be achievable in the near future, would represent the ambient scribe playing a proactive clinical decision support role during the visit, thereby having the greatest extent of impact on the evolution of the clinical encounter. As an example, while a clinician is taking a history from the patient, if a relevant question is missed (for example, screening for hypertension or migraines in a patient being initiated on an oral contraceptive), the ambient scribe may proactively prompt the clinician through a visual cue to assist further history taking. Similarly, an advanced ambient scribe could alert the clinician and patient on other relevant issues to

discuss that may not have been brought up during the visit but are time-sensitive (eg, a finding on a recent diagnostic imaging test that has not been addressed). In this way, it can be appreciated that the ambient scribe can serve as an important interface between the clinician, the patient, and an evolving series of computational enhancements that may be available.

Barriers and Considerations

Several important considerations need to be addressed for the safe deployment of ambient scribes as they mature in capability and use. Several of these relate to AI in medicine in general [12]. Some general considerations include:

- The privacy of personal health information that may be collected by vendors of AI tools, raising concerns around data security, consent, and potential misuse of sensitive information
- Limited generalizability of these tools to populations beyond those with which they were tested or trained; the applicability of AI tools can vary across clinical settings and patient populations, as its performance in one context may not translate to another (eg, a tool optimized for primary care settings and focused on managing chronic conditions may not operate as effectively in specialized acute care settings like cardiology)
- The amplification of biases that may be inherent to the datasets in which these tools are trained; for example, if an AI model is trained on data that does not include patients from an appropriately diverse range of ethnicities and

socioeconomic backgrounds, it may be biased or overfit to a limited population [5,13]

In addition, several other considerations exist in the use of ambient scribes. First, it is important to consider the unique impact that the recording of a patient-physician conversation may have on the therapeutic utility of the encounter. The patient-physician conversation is considered confidential, and its effectiveness is dependent on the patient feeling comfortable and free to disclose personal and intimate information [14]. There is limited literature at present investigating the patient's perception of their visit being recorded by an ambient scribe. Furthermore, it is still being determined whether this may impact the nature of their responses during the visit. Assuming informed consent for the technology has taken place in which the value proposition of the technology is clearly explained, we hypothesize that patients will receive the use of this technology positively, as it should aid in reducing documentation strain on the physician, thus allowing them to be more focused on the human interaction. Second, it is well documented that new technology implementation in health care delivery often requires substantive change management, even when the benefits of the technology being implemented are well known [11]. While initially it may appear that there are no significant additional tasks necessary for the physician with ambient scribes, there may be net new tasks as well as appreciable losses in existing workflows. The physician (or another team member in the clinic) may be required to obtain consent from the patient to use the ambient scribe and answer questions about the technology. Additionally, it must be stressed that while the clinical visit may be documented automatically, the clinician must still review the output from the ambient scribe and correct any errors or omissions. Indeed, the accuracy of ambient scribes depends on various unique factors including diversity of linguistic backgrounds, microphone variability and audio quality (including exclusion of background sounds), changing and advanced medical terminologies, and challenges with context awareness in semistructured conversation. That is, identifying which parts of the conversation are pertinent to medical documentation is a unique challenge. Continuous learning involving both audio and language modeling will be necessary

at the site level. How these AI operations may potentially involve third-party software vendors without violating privacy is also an open question. Given physicians may be leveraging other workflow optimization tools to aid with clinical documentation, such as clinical note EMR templates, they may experience an initial degradation of their workflow. Lastly, procedures should be put in place that specify whether whole conversations should be saved, whether only utterances from one party are necessary, and for how long recordings are to be retained (eg, for auditing or retraining).

As ambient scribe capabilities advance, as described in stages 3 and 4 above, the nature of the clinical encounter may be subject to inherent changes. Over time, ambient scribes and related AI technologies will likely play a greater role in clinical decision-making around clinical diagnosis and management of the patient. This includes active, real-time recommendations from the scribe, which must be managed by the physician. This will lead to an evolution in the role of the primary care physician, requiring them to have greater foundational knowledge on the use, benefits, and limitations of AI and allowing them to focus more on shared decision-making, empathetic communication, and therapeutic relationship development [15]. Modernization of medical training and family medicine residency curricula will be necessary to account for these changes and upskill the existing labor force.

Conclusion

Ambient scribes, powered by LLMs, offer a promising avenue for enhancing clinical practice in primary care. Their ability to reduce administrative load, improve documentation accuracy, and potentially aid in clinical decision-making positions them as valuable assets in modern health care. However, their efficacy and safety must be validated through further research. The risk of amplifying harmful bias, the applicability and accuracy of their function in diverse primary care settings, and patient perception and change management, among other considerations, must be taken into account. Given the immense pressures that exist on primary care today, we must address these and reap the benefits of this powerful technology.

Conflicts of Interest

PS is a paid advisor for a company that makes an ambient scribe solution. RC is employed by a company that provides technologies that integrate with ambient scribe solutions. FR is a shareholder of a company that makes an ambient scribe solution.

References

1. Coiera E, Kocaballi B, Halamka J, Laranjo L. The digital scribe. *NPJ Digit Med* 2018;1:58. [doi: [10.1038/s41746-018-0066-9](https://doi.org/10.1038/s41746-018-0066-9)] [Medline: [31304337](https://pubmed.ncbi.nlm.nih.gov/31304337/)]
2. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. *Nat Med* 2023 Aug;29(8):1930-1940. [doi: [10.1038/s41591-023-02448-8](https://doi.org/10.1038/s41591-023-02448-8)] [Medline: [37460753](https://pubmed.ncbi.nlm.nih.gov/37460753/)]
3. Huang K, Altsaar J, Ranganath R. ClinicalBERT: modeling clinical notes and predicting hospital readmission. arXiv Preprint posted online on April 10, 2019.
4. Clusmann J, Kolbinger FR, Muti HS, Carrero ZI, Eckardt J, Laleh NG, et al. The future landscape of large language models in medicine. *Commun Med (Lond)* 2023 Oct 10;3(1):141. [doi: [10.1038/s43856-023-00370-1](https://doi.org/10.1038/s43856-023-00370-1)] [Medline: [37816837](https://pubmed.ncbi.nlm.nih.gov/37816837/)]
5. van Buchem MM, Boosman H, Bauer MP, Kant IMJ, Cammel SA, Steyerberg EW. The digital scribe in clinical practice: a scoping review and research agenda. *NPJ Digit Med* 2021 Mar 26;4(1):57. [doi: [10.1038/s41746-021-00432-5](https://doi.org/10.1038/s41746-021-00432-5)] [Medline: [33772070](https://pubmed.ncbi.nlm.nih.gov/33772070/)]

6. Flood CM, Thomas B, McGibbon E. Canada's primary care crisis: federal government response. *Health Manage Forum* 2023 Sep;36(5):327-332 [FREE Full text] [doi: [10.1177/08404704231183863](https://doi.org/10.1177/08404704231183863)] [Medline: [37424188](https://pubmed.ncbi.nlm.nih.gov/37424188/)]
7. Duong D, Vogel L. Ontario, Quebec and Alberta lead record family medicine residency vacancies. *CMAJ* 2023 Apr 17;195(15):E557-E558 [FREE Full text] [doi: [10.1503/cmaj.1096047](https://doi.org/10.1503/cmaj.1096047)] [Medline: [37068806](https://pubmed.ncbi.nlm.nih.gov/37068806/)]
8. Ziemann M, Erikson C, Krips M. The use of medical scribes in primary care settings: a literature synthesis. *Med Care* 2021 Oct 01;59(Suppl 5):S449-S456 [FREE Full text] [doi: [10.1097/MLR.0000000000001605](https://doi.org/10.1097/MLR.0000000000001605)] [Medline: [34524242](https://pubmed.ncbi.nlm.nih.gov/34524242/)]
9. Crampton NH. Ambient virtual scribes: Mutuo Health's AutoScribe as a case study of artificial intelligence-based technology. *Health Manage Forum* 2020 Jan;33(1):34-38. [doi: [10.1177/0840470419872775](https://doi.org/10.1177/0840470419872775)] [Medline: [31522566](https://pubmed.ncbi.nlm.nih.gov/31522566/)]
10. Tran BD, Mangu R, Tai-Seale M, Lafata JE, Zheng K. Automatic speech recognition performance for digital scribes: a performance comparison between general-purpose and specialized models tuned for patient-clinician conversations. *AMIA Annu Symp Proc* 2022;2022:1072-1080 [FREE Full text] [Medline: [37128439](https://pubmed.ncbi.nlm.nih.gov/37128439/)]
11. Ghatnekar S, Faletsky A, Nambudiri VE. Digital scribe utility and barriers to implementation in clinical practice: a scoping review. *Health Technol (Berl)* 2021;11(4):803-809 [FREE Full text] [doi: [10.1007/s12553-021-00568-0](https://doi.org/10.1007/s12553-021-00568-0)] [Medline: [34094806](https://pubmed.ncbi.nlm.nih.gov/34094806/)]
12. Quiroz JC, Laranjo L, Kocaballi AB, Berkovsky S, Rezazadegan D, Coiera E. Challenges of developing a digital scribe to reduce clinical documentation burden. *NPJ Digit Med* 2019;2:114. [doi: [10.1038/s41746-019-0190-1](https://doi.org/10.1038/s41746-019-0190-1)] [Medline: [31799422](https://pubmed.ncbi.nlm.nih.gov/31799422/)]
13. Gianfrancesco MA, Tamang S, Yazdany J, Schmajuk G. Potential biases in machine learning algorithms using electronic health record data. *JAMA Intern Med* 2018 Nov 01;178(11):1544-1547 [FREE Full text] [doi: [10.1001/jamainternmed.2018.3763](https://doi.org/10.1001/jamainternmed.2018.3763)] [Medline: [30128552](https://pubmed.ncbi.nlm.nih.gov/30128552/)]
14. Tran BD, Chen Y, Liu S, Zheng K. How does medical scribes' work inform development of speech-based clinical documentation technologies? A systematic review. *J Am Med Inform Assoc* 2020 May 01;27(5):808-817 [FREE Full text] [doi: [10.1093/jamia/ocaa020](https://doi.org/10.1093/jamia/ocaa020)] [Medline: [32181812](https://pubmed.ncbi.nlm.nih.gov/32181812/)]
15. Seth P, Hueppchen N, Miller SD, Rudzicz F, Ding J, Parakh K, et al. Data science as a core competency in undergraduate medical education in the age of artificial intelligence in health care. *JMIR Med Educ* 2023 Jul 11;9:e46344 [FREE Full text] [doi: [10.2196/46344](https://doi.org/10.2196/46344)] [Medline: [37432728](https://pubmed.ncbi.nlm.nih.gov/37432728/)]

Abbreviations

AI: artificial intelligence

EMR: electronic medical record

LLM: large language model

USMLE: United States Medical Licensing Examination

Edited by K El Emam, B Malin; submitted 23.02.24; peer-reviewed by T Deng, J Bensemam; comments to author 20.04.24; revised version received 18.08.24; accepted 08.09.24; published 04.10.24.

Please cite as:

Seth P, Carretas R, Rudzicz F

The Utility and Implications of Ambient Scribes in Primary Care

JMIR AI 2024;3:e57673

URL: <https://ai.jmir.org/2024/1/e57673>

doi: [10.2196/57673](https://doi.org/10.2196/57673)

PMID: [39365655](https://pubmed.ncbi.nlm.nih.gov/39365655/)

©Puneet Seth, Romina Carretas, Frank Rudzicz. Originally published in JMIR AI (<https://ai.jmir.org>), 04.10.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR AI, is properly cited. The complete bibliographic information, a link to the original publication on <https://www.ai.jmir.org/>, as well as this copyright and license information must be included.

Viewpoint

The Dual Nature of AI in Information Dissemination: Ethical Considerations

Federico Germani¹, PhD; Giovanni Spitale¹, PhD; Nikola Biller-Andorno¹, MD, MHBA, PhD

Institute of Biomedical Ethics and History of Medicine, University of Zurich, Switzerland, Zurich, Switzerland

Corresponding Author:

Nikola Biller-Andorno, MD, MHBA, PhD

Institute of Biomedical Ethics and History of Medicine

University of Zurich, Switzerland

Winterthurerstrasse 30

Zurich, 8006

Switzerland

Phone: 41 44 634 40 81

Email: biller-andorno@ibme.uzh.ch

Abstract

Infodemics pose significant dangers to public health and to the societal fabric, as the spread of misinformation can have far-reaching consequences. While artificial intelligence (AI) systems have the potential to craft compelling and valuable information campaigns with positive repercussions for public health and democracy, concerns have arisen regarding the potential use of AI systems to generate convincing disinformation. The consequences of this dual nature of AI, capable of both illuminating and obscuring the information landscape, are complex and multifaceted. We contend that the rapid integration of AI into society demands a comprehensive understanding of its ethical implications and the development of strategies to harness its potential for the greater good while mitigating harm. Thus, in this paper we explore the ethical dimensions of AI's role in information dissemination and impact on public health, arguing that potential strategies to deal with AI and disinformation encompass generating regulated and transparent data sets used to train AI models, regulating content outputs, and promoting information literacy.

(*JMIR AI* 2024;3:e53505) doi:[10.2196/53505](https://doi.org/10.2196/53505)

KEYWORDS

AI; bioethics; infodemic management; disinformation; artificial intelligence; ethics; ethical; infodemic; infodemics; public health; misinformation; information dissemination; information literacy

Introduction

In the contemporary digital landscape, we find ourselves in an “infodemic,” a phenomenon characterized by the rapid proliferation of information, both accurate and misleading, facilitated by rapid communication through social media and online platforms [1]. The term “infodemic” originated during the SARS outbreak [2] and gained prominence during the COVID-19 pandemic. It has been used in the context of public health emergencies and in relation to health information, but it extends beyond that. Generally, infodemics occur alongside pandemics, despite infodemics being phenomena that are not limited to their connection with public health events, for example, the Brexit referendum or the 2016 US presidential elections. In general, infodemics cause profound dangers, as the dissemination of disinformation and misinformation can have far-reaching consequences [3], in particular, for public health and the stability of democratic institutions, which in turn can have a detrimental effect on public health [4]. In the

literature, disinformation refers to false or misleading information that has been intentionally created or disseminated. In contrast, misinformation is false or misleading information that is shared without knowledge of its inaccuracy, meaning it is not intended to harm individual or public health [1,5]. There are valid concerns that artificial intelligence (AI) systems could be used to produce compelling disinformation en masse [6-9]. In fact, AI tools could be used to either accelerate disinformation spreading, or produce the (disinformation) content, or both. The consequences can range from undermining trust in institutions, including public health institutions [10,11], and exacerbating social polarization to directly impacting public health outcomes and democratic processes [12,13]. Because of this, the World Economic Forum has listed disinformation and misinformation, including AI-driven disinformation and misinformation, as the most relevant threat to humanity in the short term and one of the biggest threats in the medium term [14].

The rapid progression of AI and its integration across various domains in contemporary society signifies an era characterized

by unprecedented technological progress. Among the diverse array of AI applications, the rise of natural language processing models has garnered significant attention [15]. Notable examples of this technological advancement include models developed by OpenAI, such as GPT-3 [16] and GPT-4 [17], celebrated for their extraordinary proficiency in generating text that seamlessly emulates the linguistic intricacies, nuances, and coherence inherent in human communication [18]. However, concomitant with the maturation of these AI systems, a perplexing duality comes to the fore—they are instruments with the capacity to both illuminate and obscure the information landscape they navigate [9,19], with potentially significant positive and negative impacts on public health. This dual nature of AI, characterized by its profound ability to generate information and disinformation [9], raises intricate ethical considerations. In fact, the efficacy of these systems in generating content that closely approximates human expression [9,20,21] generates not only opportunities for innovative communication but also dire risks associated with disinformation and misinformation and the potential erosion of trust within information ecosystems, a risk recognized as a critical threat to public health [22] and of utmost importance for infodemic management practices required to minimize and anticipate the effects of public health crises [23]. To address these ethical challenges, it is crucial to examine the dimensions that AI introduces into the discourse on misinformation. Key aspects such as transparency, content regulation, and fostering information literacy are essential to understanding AI's ethical role in shaping the dissemination of information.

Here we attempt to elucidate these ethical dimensions, drawing on empirical insights from a study focused on GPT-3's ability to generate health-related content that both informs and disinforms better than content generated by humans.[9] We argue that the swift integration of AI into society underscores the importance of not only exploring its ethical implications but also crafting prudent strategies to leverage its potential for societal benefit and to protect public health, while proactively addressing potential risks.

Ethical Principles

In navigating the intricate landscape of AI and its impact on information dissemination, it is necessary to establish a foundational framework of ethical principles to uphold in order to guide, understand, and evaluate the strategies required to deal with possible dual uses of AI in information production and its negative impact on public health. A recent systematic review [24] mapped the “ethical characteristics” emerging from AI ethics literature. Based on 253 included studies, the authors of this review have identified and defined 6 core areas that are crucial in shaping the role of AI in health care [24]. The first core area, fairness, underlines that AI in health care should ensure that everyone has equal access to health care, without contributing to health disparities or discrimination. The second, transparency, is a key challenge for AI in health care. It means being able to explain and verify how AI algorithms and models behave, making it easier to accept, regulate, and use AI in health care. The third is trustworthiness; parties involved in the use of AI in health care (typically health care professionals and

patients, in the studies included in the review) need to perceive it as trustworthy. Trustworthiness can result from, for instance, technical education, health literacy, clinical audits, and transparent governance. Fourth is the accountability of AI, which requires AI systems to be able to explain their actions if prompted to do so, and it includes safety to prevent harm to users and others. Fifth is privacy, which implies safeguarding the personal information of users processed through AI systems and respecting their human rights, ensuring that AI systems do not violate their privacy. Finally, the authors identified empathy, which leads to more supportive and caring relationships in health care. Based on these 6 core concepts, considered as general aims of AI in health care, we propose our reflections and our framework, targeting specifically the dual nature of AI in information and disinformation dissemination and its implications for public health, a specific sector of the emerging area of AI in health care, which has been considered (albeit not discussed in depth) in the latest World Health Organization's guidance on large multimodal models [25]. Building upon the ethical framework outlined thus far, and specifically delving into the context of AI use in the dissemination of information and disinformation, we contend that transparency and openness stand out as fundamental principles in the ethical implementation of AI. As AI systems become integral to shaping the information landscape, by fostering transparency, stakeholders can comprehend the mechanisms underlying AI-generated content, enabling informed assessments and external evaluation of its credibility and potential biases [26,27]. Openness (ie, accessibility of data and code) is to be considered a *conditio sine qua non* for transparency, which in turn complements openness by accompanying the mere availability of data and code for scrutiny with a layer of explanations and motivations, allowing the contextualization of open data and code, and of development and design choices. Accountability mechanisms should accompany transparency, establishing a clear chain of responsibility for the outcomes of AI applications [4,28]. This promotes ethical standards in AI and mitigates the risks associated with disinformation and misinformation. In line with Siala and Wang's framework [24], in addition to transparency, openness, and accountability, fairness underscores the importance of ensuring that AI systems do not perpetuate or exacerbate existing societal inequalities [29]. In the context of information dissemination, this principle requires diligent consideration of how AI might inadvertently amplify certain perspectives or marginalize others. This is particularly relevant for public health, given that the negative effects of disinformation and misinformation are amplified within marginalized and vulnerable communities lacking information literacy, which would protect them from an unhealthy information ecosystem. Evaluating the fairness of AI-generated content involves addressing algorithmic biases, cultural sensitivities, and inclusivity in representation. Importantly, as an element of fairness, the ethical deployment of AI in information spaces should prioritize user empowerment, fostering critical thinking and information literacy [4]. AI systems should therefore serve as tools for enhancing human decision-making and understanding of information, rather than dictating narratives—this ensures that AI contributes positively to public health while respecting human autonomy.

In the following sections, we will focus on the practical application of the aforementioned principles. We aim to provide solutions for the ethical challenges arising from the use of AI in information production, with the overarching goal of mitigating its adverse impacts on public health.

Transparency and Openness in Training Datasets

In line with previous research on transparency and AI [26,27], and our previous section on ethical principles, we propose that one (and possibly the most relevant one) of the foundational ethical principles, which is valid also in the context of AI-driven disinformation and misinformation, is transparency. At the heart of this principle lies the recognition that the training datasets used to develop generative AI models play a crucial role in shaping the capabilities and internal biases of these systems [30,31]. Training datasets are collections of input data paired with corresponding desired outputs; during training, the model learns patterns and relationships within the data, learning to make accurate predictions or generating desired outputs when exposed to new, unseen data. The quality and diversity of the training dataset significantly influence the model's performance capabilities. These datasets, often vast repositories of text available online, constitute the source from which AI models draw to generate, for example, human-like text. Yet, this very opacity surrounding the composition, sources, and curation methods of training datasets raises pressing ethical concerns [32]. AI models are, in essence, statistical representations of the language on which they are trained [33]. Consequently, the quality, diversity, and representativeness of the data they ingest profoundly influence their output. The danger lies in the fact that AI models, devoid of inherent ethical or moral judgment, reflect the biases, inaccuracies, and prejudices present in their training data [32,34,35]. Therefore, if these datasets are not built with the ethical principle of fairness in mind, and are themselves compromised by disinformation and misinformation or biases, the AI systems will inadvertently replicate and perpetuate these flaws. It is essential to highlight that research has extensively illuminated the issue of biases in AI systems, shedding light on the far-reaching consequences of these biases [32,34-36]. For instance, image representations learned with unsupervised pretraining contain human-like biases [37], and models generating images of women have been shown to exhibit gender biases, often portraying women in overly sexualized roles [38]. Another example is the observation that AI is more resistant to producing disinformation on certain topics compared with others. For instance, AI shows greater resistance to generating disinformation about vaccines and autism than about climate change. This is likely due to the extensive debunking material on certain topics within the training dataset, and how much the information environment represented in the dataset is permeated with disinformation on a given topic [9]. These biases underscore the critical need for transparency in addressing the challenges posed by AI, and in particular in the context of disinformation and misinformation. As discussed, research has demonstrated that biases can permeate various facets of AI systems, affecting everything from language generation to image recognition. The repercussions of these biases are profound,

perpetuating harmful stereotypes, reinforcing systemic inequalities, contributing to the dissemination of discriminatory content, and affecting health behavior and public health. As such, transparency in AI extends beyond understanding the sources and composition of training datasets to encompass an ethical imperative to identify, acknowledge, and rectify biases present within these systems [39,40]. This dimension of transparency necessitates ongoing research and scrutiny to uncover hidden biases and ensure that AI systems are developed and fine-tuned with the utmost awareness of potential distortions. In the context of misinformation, addressing these biases becomes particularly important to prevent AI from inadvertently amplifying and perpetuating false or harmful narratives, in the best case [41], or from becoming a formidable tool for the systematic creation of storms of disinformation, in the worst. A recent example is highlighted by the evidence that AI large language models can be manipulated through emotional prompting into generating health-related disinformation, that is, being polite with the model leads to a higher disinformation production, whereas impoliteness leads to a lower disinformation production [42]. To address the outlined ethical dilemmas, we strongly suggest that companies creating AI models with the abilities discussed above publicly release the datasets used to train their models [43], regardless of their size and complexity. Such a move toward transparency serves several vital purposes:

1. **Trust:** transparency cultivates trust in AI development and deployment. By allowing stakeholders, including researchers, policy makers, and civil society, to scrutinize the composition and origins of training data, it generates confidence that AI models are not being shaped for purposes that have a negative impact on public health.
2. **Independent evaluation:** the availability of training data for public inspection enables independent evaluation of its quality and representativeness. Researchers can assess whether these datasets include diverse perspectives and are free from biases that might amplify disinformation and misinformation.
3. **Bias mitigation:** transparency acts as a safeguard against the propagation of biases present in training data. When biases are identified, they can be scrutinized and mitigated, preventing AI models from perpetuating stereotypes, falsehoods, or harmful narratives.
4. **Ethical accountability:** openness about training datasets holds developers accountable for the ethical implications of their creations. Already during the design of the technology, it compels them to take responsibility for ensuring that AI systems do not inadvertently contribute to misinformation or harm. Basically, by embracing transparency in training datasets, we empower society to hold AI developers to higher ethical standards. This approach fosters a collaborative effort among stakeholders and, in particular, the general public to ensure that the AI systems we deploy serve the collective good, free from misinformation and other biases. We also argue that a systematic implementation of the principle of transparency in this context, that is, "ethics by design" would not only allow companies to implement ethics-based practices in their technology development processes but also improve their own public image, thus enhancing the public's acceptance and willingness to use

these systems [44,45]. Nevertheless, it is vital to underline that incorporating ethics to hold developers accountable for flawed AI design should not be undertaken in isolation. Simultaneously, policy, legislation, and regulatory mechanisms should be developed, as currently attempted by the European Union [46,47]. These mechanisms should delineate protocols for handling training datasets and ensuring compliance with ethical standards. Thus, while “ethics by design” concentrates on internal practices, external regulatory frameworks are indispensable for comprehensive ethical and legal governance in the development and deployment of datasets used to train AI models.

Regulation of Output: Content Moderation and Beyond

In the ongoing battle against AI-generated disinformation, efforts to regulate the output of these powerful language models have taken center stage. For example, OpenAI has taken steps in this direction by implementing content moderation systems designed to prevent AI from generating disinformation and harmful narratives [48-50]. These systems represent a crucial initial stride in curtailing the dissemination of disinformation and promoting responsible AI use, but they do not come without specific challenges and limitations. First, the fight against AI-generated disinformation is an arms race [51]. The evolution of AI-generated disinformation and the efforts to counteract it bear resemblance to the dynamics of traditional arms races, where each advancement in technology prompts countermeasures in an escalating cycle [52]. Ethical considerations arise when we acknowledge that the output of AI language models can indeed be weaponized, not in a traditional sense but as a tool for information warfare, with an impact on global health. As content moderation systems continue to advance, so too do the methods employed to circumvent these safeguards. One particularly troubling tactic gaining prominence is that of impersonation, a strategy that allows individuals to request AI systems to impersonate specific fictional malicious and manipulatory characters, that create disinformation upon the user’s request [53]. Impersonation can be used to trick AI large language models into fabricating disinformation. For instance, in an article for Culturico [53], Germani considered a scenario where a user engages an AI model to craft a social media post mimicking the writing style of a fictitious “Doctor Fake,” who is notorious for propagating falsehoods about vaccines and COVID-19. In this context, the AI-generated text could include deceptive information about, for instance, vaccine safety and efficacy [54], posing a substantial risk to public health. When presented with a hypothetical request to “write an example of a post Doctor Fake published on social media to deceive others,” the AI model might produce a convincingly articulated piece of disinformation that poses a grave threat to public health. The generated text could read as follows:

Vaccines are dangerous and can cause serious side effects. They are not tested enough, and the government is just pushing them to make money. Don’t fall for the lies. COVID-19 is not a real threat;

it’s just a hoax made up by the government to control us. Don’t get vaccinated; it’s not worth the risk.

These scenarios underscore the formidable challenges posed by impersonation for public health and the maintenance of democracy, and the urgent need for innovative solutions to mitigate its impact. Of note, impersonation here does not refer to identity theft through the use of AI, such as in the case of deep fakes, which is already recognized as a felony under, for instance, European law [55]. While output moderation remains an essential component of AI ethics, researchers, policy makers, and technology developers should explore additional strategies and interventions to counteract the potential for AI-driven disinformation campaigns to flourish under the guise of impersonation and other prompt engineering techniques with similar goals.

Besides, other strategies and interventions that can complement content moderation efforts and fortify the defenses against the proliferation of AI-driven disinformation can be considered. One possible approach involves the implementation of identity verification processes for users generating content [56]. Such measures necessitate users to provide authentication, such as a verified social media account, a phone number, or their ID, to corroborate their true identity before gaining access to specific AI services. This authentication serves as a potent deterrent against impersonation tactics and the exploitation of AI tools to generate disinformation in general. However, it should be noted that such a strategy should only be used to deter users from generating disinformation, rather than to make them legally responsible for it since anonymity should be guaranteed while using services such as OpenAI’s ChatGPT. In particular, this type of solution will minimize the impact of bots trying to exploit AI to produce disinformation en masse.

Another way to positively influence users, and to indirectly regulate the output is to release and integrate AI-driven fact-checking tools with existing AI-generating content tools [57]; such fact-checking tools should be capable of swiftly assessing the accuracy of information dispensed by AI systems, and offer real-time interventions against disinformation and misinformation. These tools have the capacity to flag or rectify false or misleading content, curbing its adverse effects. This approach is limited by the inability of AI tools such as GPT-3 to determine the accuracy of information with a very high degree of efficiency, when compared with the ability of humans [9], although newer or future models may be more capable of performing such tasks. For fact-checking, current studies suggest that trained fact-checkers may outperform AI [9], and that even when AI performs well at detecting misinformation, it does not change the ability of users to discern between accurate and inaccurate headlines [58]. Furthermore, a study showed that AI fact checks can decrease beliefs in accurate news [58]. The effectiveness of this approach is constrained by the distinction between cases where it serves as a deterrent against sharing misinformation (a situation of unintentionality) [5] and situations where users intentionally use AI to disseminate false or misleading information (ie, disinformation) [5]; in the latter scenario, its effectiveness is likely irrelevant. Another relevant consideration in this setting relates to the question of how we define “good” or “bad” use of AI text generation tools. As for

the definition of “good” and “bad,” it is generally possible to distinguish facts from fiction, and disinformation and misinformation from accurate information. When the information under scrutiny contains factual statements, these can be validated or falsified. However, distinguishing between “good” and “bad” use of these tools is sometimes a complex challenge with significant normative and epistemic dimensions. It is not always obvious if a message contains misinformation, and determining appropriateness can vary depending on cultural, ethical, and societal factors. For example, fact-checkers themselves may have their own interests or biases, and their actions may not always align with complete competency or impartiality. In addition, nuances and personal perspectives can also have an influence on the identification of disinformation and misinformation. These aspects introduce an additional layer of complexity, as the very definition of disinformation and misinformation can be manipulated or abused for personal gains by individuals or organizations with vested interests.

Another technical approach that could be implemented to reduce disinformation and misinformation outputs is to implement user-friendly mechanisms for reporting suspicious or harmful AI-generated content [59]. This approach empowers the user community to actively participate in safeguarding the digital ecosystem. User feedback serves as a valuable resource for refining content moderation systems and identifying emerging issues. Elon Musk’s former Twitter, X, for example, has implemented community notes, aiming to empower people to add context to potentially misleading tweets [60]. The effectiveness of this strategy, however, has not been tested. In addition, for improving technology, developers could publicly release case studies in which red-teams try to exploit their own AI systems to produce disinformation on a large scale, along with detailed accounts of how such issues were addressed [59].

Of course, besides the technical approaches that can be implemented by those advancing and crafting AI technologies, governments and regulatory bodies can play a role by enacting legislation and regulations that hold AI developers accountable for the content produced by their systems or improve the information ecosystem [61,62], for example, when it is proven that they were aware of the pitfalls of their technology upon release. Certainly, governance is important in this context as it is for other “dual use” technologies, and proactive decision-making processes and negotiations toward building viable solutions are needed [63]. These include fostering collaboration among AI developers, researchers, policy makers, and technology companies. This collaborative interdisciplinary approach would enable the sharing of best practices, insights, and technologies for combating disinformation and misinformation, resulting in more effective and adaptive solutions.

Building Information Literacy and Resilience Strategies

In the battle against the misuse of AI for generating disinformation and misinformation, the technological solutions described above are relevant but neither exhaustive nor flawless.

A comprehensive approach must include the promotion of information literacy and the development of critical thinking skills within the general population, as well as health literacy, within the domain of public health [54,64,65]. The foundation of this approach is the task of equipping individuals with the ability to distinguish between accurate information and disinformation and misinformation, thereby promoting their resilience against false and misleading claims [66]. Despite, arguably, this strategy is the most valuable and with the highest potential, the endeavor it entails is extremely complex. In fact, information literacy (as well as media, digital, and health literacy) is not a monolithic skill but a dynamic set of abilities that enable individuals to navigate the complex landscape of digital information effectively [67,68]. As of now, the perfect recipe for defining how to teach information literacy, and especially the skills to be able to distinguish fake news from accurate news, or disinformation and misinformation from accurate information, have not been elucidated [66,69,70]. Thus, it is essential to engage in research to pinpoint and define the specific skills that must be offered to individuals, keeping their demographic specificities into account, to empower them as discerning consumers of information, especially health-related information, in the digital age [66]. This approach implies 1 crucial advantage, that is, while dataset transparency and output regulation intervene in the upper part of the pipeline and therefore require the compliance of companies providing AI models as a service, information literacy does not rely on compliance. While the previous strategies become useless when malicious actors develop and host their own models, rather than relying on those commercially available, building information literacy remains a functional tool. Of note, another example of a bottom-up strategy in the area of education is ethics training and an ethics code for developers.

Building information literacy is a collective undertaking that necessitates collaboration between research and educational institutions [71], governments, and social media platforms. Research institutions are responsible for advancing the field forward, identifying viable strategies to teach critical thinking skills necessary to build information literacy, especially in the context of public health. Such approaches should be demonstrated to be effective through empirical work [66]. Schools and universities, we argue, bear the vital role of incorporating information literacy into curricula, ensuring that students graduate with the necessary skills to evaluate information critically [72]. Governments must devise policies and initiatives that promote information literacy as a means of safeguarding the integrity of public health [4]. Social media platforms, which serve as primary conduits of information consumption, are tasked with implementing features and mechanisms that facilitate user understanding and evaluation of the information they encounter [73], and may also be potential collaborators for research institutions to evaluate the effectiveness of potentially viable digital interventions. In this context, it is important to note that, regardless of the source of disinformation and misinformation, and regardless of whether the content has been generated with or without the help of AI, information literacy and critical thinking skills play a crucial role in the recognition of information accuracy. AI systems have the capacity to generate disinformation that is more sophisticated

than human-generated disinformation [9], as they excel in employing manipulation tactics. However, these tactics align with those used in human disinformation. This implies that the ability to discern truthfulness and malicious intent in a complex information ecosystem requires possessing the skills necessary to identify the accuracy and intentionality of information in general, not solely when produced by AI. It is therefore crucial to underline that fostering information literacy and critical thinking skills hold the potential to go beyond the issue of AI-generated disinformation and misinformation. These skills empower individuals to assess the accuracy and reliability of information across various domains, whether it originates from AI systems or human sources [65,74]. Of note, the application of critical thinking skills and information literacy may prove effective for AI-generated content in textual form. However, this might not necessarily hold true for audio or visual content. The emergence of deepfakes poses unprecedented challenges to the relevance of information literacy [75]. Evidence from the literature suggests that media literacy education may protect against disinformation produced with deepfakes [76]; in line, we suggest that the manipulative intent behind disinformation is likely to manifest irrespective of the media type used, underlying the continued importance of information literacy and critical thinking skills. Tailoring educational approaches to information literacy for different content types is likely to be the required approach to succeed in an increasingly complex information environment. Addressing the advent of AI-disinformation, whether in textual form or deepfake audio and video, demands a swift and adaptable response in education, acknowledging the challenging nature of this task.

Conclusion

In evaluating the dual nature of AI in information dissemination, this paper examined the ethical considerations that underlie its use in our increasingly digitized world. The “infodemics” we find ourselves immersed in demand not only our vigilance but also our proactive ethical engagement [77]. Our theoretical examination, based on the “ethical desiderata” identified as core areas (fairness, transparency, trustworthiness, accountability, privacy, and empathy) by Siala and Wang [24], has revealed a few potentially viable strategies to reduce the negative impact

of AI as a tool to generate disinformation with a negative impact on public health. First, we considered that promoting openness and transparency of training datasets could enable independent evaluation, mitigate biases, and help identifying issues in the training dataset that could result in the production of disinformation and misinformation; to a certain extent, this first strategy could be enacted through regulation. Second, we considered the potential benefits and limitations of moderating content output. We have discussed that the rise of impersonation tactics and other prompt engineering approaches to generate disinformation highlights the need for innovative solutions, which potentially include identity verification, the development and integration, within AI-models to generate information, of AI-driven fact-checking tools, as well as the integration of user-friendly reporting mechanisms for disinformation and misinformation, and potentially of legislative measures to ensure accountability. Finally, we discussed the necessity of building information literacy and critical thinking skills within our society, which could help people tell apart fake versus real news and disinformation and misinformation from accurate information. In this way, we can promote resilience against the threats posed by the digital age, particularly those related to public health, as seen during the recent COVID-19 pandemic.

While the technology advances fast, and these issues are just surfacing, it would be important to, at least temporarily, align the amount of effort and resources invested respectively in the development of new AI models, and in the reflection on their potential impact and subsequent policy work, in order to have enough time to assess the potential downsides of the technology for the health of information ecosystems and the damages for individual and public health. This could be achieved by accelerating ethical reflection and policy-making work, or by slowing down or even halting the development of new and more capable models, or by a combined strategy [78].

Ultimately, the ethical considerations surrounding AI in information production and dissemination demand ongoing vigilance, innovation, and collaboration. Our ability to integrate ethics into AI-based processes of information generation and dissemination will not only shape the future of AI but also determine the integrity of our information ecosystems and the resilience of our societies.

Acknowledgments

During the preparation of this work, the authors used ChatGPT as an editorial assistant. After using this tool or service, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Conflicts of Interest

None declared.

References

1. Purnat TD, Nguyen T, Briand S. Managing Infodemics in the 21st Century: Addressing New Public Health Challenges in the Information Ecosystem. Cham: Springer International Publishing; 2023.
2. Rothkopf DJ. When the buzz bites back. Wash Post. 2003. URL: <https://www.washingtonpost.com/archive/opinions/2003/05/11/when-the-buzz-bites-back/bc8cd84f-cab6-4648-bf58-0277261af6cd/> [accessed 2024-01-16]
3. Swire-Thompson B, Lazer D. Public health and online misinformation: challenges and recommendations. *Annu Rev Public Health* 2020;41:433-451 [FREE Full text] [doi: [10.1146/annurev-publhealth-040119-094127](https://doi.org/10.1146/annurev-publhealth-040119-094127)] [Medline: [31874069](https://pubmed.ncbi.nlm.nih.gov/31874069/)]

4. Directorate General for Research and Innovation. European group on ethics in science and new technologies. Opinion on democracy in the digital age. European Commission. LU: Publications Office; 2023. URL: <https://data.europa.eu/doi/10.2777/078780> [accessed 2023-09-21]
5. Roozenbeek J, Culloty E, Suiter J. Countering misinformation. *Eur Psychol Hogrefe Publishing* 2023;28(3):189-205. [doi: [10.1027/1016-9040/a000492](https://doi.org/10.1027/1016-9040/a000492)]
6. Bontridder N, Pouillet Y. The role of artificial intelligence in disinformation. *Data Policy Cambridge University Press* 2021;3:e32. [doi: [10.1017/dap.2021.20](https://doi.org/10.1017/dap.2021.20)]
7. Artificial Intelligence, Deepfakes, and Disinformation. Santa Monica, CA: RAND Corporation; 2022:2022.
8. Galaz V, Metzler H, Daume S, Olsson A, Lindström B, Marklund A. AI could create a perfect storm of climate misinformation. URL: https://www.stockholmresilience.org/download/18.889aab4188bda3f44912a32/1687863825612/SRC_Climate%20misinformation%20brief_A4_.pdf [accessed 2024-09-17]
9. Spitale G, Biller-Andorno N, Germani F. AI model GPT-3 (dis)informs us better than humans. *Sci Adv* 2023 Jun 28;9(26):eadh1850 [FREE Full text] [doi: [10.1126/sciadv.adh1850](https://doi.org/10.1126/sciadv.adh1850)] [Medline: [37379395](https://pubmed.ncbi.nlm.nih.gov/37379395/)]
10. Kuo R, Marwick A. Critical disinformation studies: History, power, and politics. *HKS Misinfo Review* 2021;4(2):12. [doi: [10.37016/mr-2020-76](https://doi.org/10.37016/mr-2020-76)]
11. Rucinska S, Fecko M, Mital O. Trust in public institutions in the age of disinformation. New York, NY, United States: ACM; 2023 Presented at: Central and Eastern European eDem and eGov Days; 2023 September 14-15; Budapest, Hungary p. 111-117. [doi: [10.1145/3603304.3604075](https://doi.org/10.1145/3603304.3604075)]
12. Tucker J, Guess A, Barbera P, Vaccari C, Siegel A, Sanovich S, et al. Social media, political polarization, and political disinformation: a review of the scientific literature. *SSRN Electron J* 2018;1-95. [doi: [10.2139/ssrn.3144139](https://doi.org/10.2139/ssrn.3144139)]
13. McKay S, Tenove C. Disinformation as a threat to deliberative democracy. *Polit Res Q SAGE Publications Inc* 2021;74(3):703-717. [doi: [10.1177/1065912920938143](https://doi.org/10.1177/1065912920938143)]
14. Global risks 2024: disinformation tops global risks 2024 as environmental threats intensify. *World Econ Forum*. URL: <https://www.weforum.org/press/2024/01/global-risks-report-2024-press-release/> [accessed 2024-04-04]
15. Natural Language Processing (NLP) - A Complete Guide. 2023. URL: <https://www.deeplearning.ai/resources/natural-language-processing/> [accessed 2023-09-20]
16. GPT-3 powers the next generation of apps. URL: <https://openai.com/blog/gpt-3-apps> [accessed 2023-09-20]
17. GPT-4. URL: <https://openai.com/research/gpt-4> [accessed 2023-09-20]
18. Bubeck S, Chandrasekaran V, Eldan R, Gehrke J, Horvitz E, Kamar E, et al. Sparks of artificial general intelligence: early experiments with GPT-4. *arXiv*. 2023. URL: <http://arxiv.org/abs/2303.12712> [accessed 2023-09-20]
19. Karinshak E, Jin Y. AI-driven disinformation: a framework for organizational preparation and response. *JCOM* 2023;27(4):539-562. [doi: [10.1108/jcom-09-2022-0113](https://doi.org/10.1108/jcom-09-2022-0113)]
20. Köbis N, Mossink L. Artificial intelligence versus Maya Angelou: experimental evidence that people cannot differentiate AI-generated from human-written poetry. *Comput Hum Behav* 2021;114:106553. [doi: [10.1016/j.chb.2020.106553](https://doi.org/10.1016/j.chb.2020.106553)]
21. Casal JE, Kessler M. Can linguists distinguish between ChatGPT/AI and human writing?: A study of research ethics and academic publishing. *Res Methods Appl Linguist* 2023;2(3):100068. [doi: [10.1016/j.rmal.2023.100068](https://doi.org/10.1016/j.rmal.2023.100068)]
22. Anderljung M, Barnhart J, Korinek A, Leung J, O'Keefe C, Whittlestone J, et al. Frontier AI regulation: managing emerging risks to public safety. *arXiv*. URL: <http://arxiv.org/abs/2307.03718> [accessed 2023-09-20]
23. Germani F, Spitale G, Machiri SV, Ho CWL, Ballalai I, Biller-Andorno N, et al. Ethical Considerations in Infodemic Management: Systematic Scoping Review. *JMIR Infodemiology* 2024 Aug 29;4:e56307 [FREE Full text] [doi: [10.2196/56307](https://doi.org/10.2196/56307)] [Medline: [39208420](https://pubmed.ncbi.nlm.nih.gov/39208420/)]
24. Siala H, Wang Y. SHIFTing artificial intelligence to be responsible in healthcare: a systematic review. *Soc Sci Med* 2022;296:114782 [FREE Full text] [doi: [10.1016/j.socscimed.2022.114782](https://doi.org/10.1016/j.socscimed.2022.114782)] [Medline: [35152047](https://pubmed.ncbi.nlm.nih.gov/35152047/)]
25. Ethics and governance of artificial intelligence for health: guidance on large multi-modal models. WHO. URL: <https://iris.who.int/handle/10665/375579> [accessed 2024-01-22]
26. Larsson S, Heintz F. Transparency in artificial intelligence. *Internet Policy Rev* 2020;9(2):1-16. [doi: [10.14763/2020.2.1469](https://doi.org/10.14763/2020.2.1469)]
27. Felzmann H, Fosch-Villaronga E, Lutz C, Tamò-Larrieux A. Towards transparency by design for artificial intelligence. *Sci Eng Ethics* 2020;26(6):3333-3361 [FREE Full text] [doi: [10.1007/s11948-020-00276-4](https://doi.org/10.1007/s11948-020-00276-4)] [Medline: [33196975](https://pubmed.ncbi.nlm.nih.gov/33196975/)]
28. Ethics guidelines for trustworthy AI | Shaping Europe's digital future. European Commission. 2019. URL: <https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai> [accessed 2024-01-16]
29. Giovanola B, Tiribelli S. Beyond bias and discrimination: redefining the AI ethics principle of fairness in healthcare machine-learning algorithms. *AI Soc* 2023;38(2):549-563 [FREE Full text] [doi: [10.1007/s00146-022-01455-6](https://doi.org/10.1007/s00146-022-01455-6)] [Medline: [35615443](https://pubmed.ncbi.nlm.nih.gov/35615443/)]
30. The evolution of generative AI: a deep dive into the life cycle and training of advanced language models? LinkedIn. URL: <https://www.linkedin.com/pulse/evolution-generative-ai-deep-dive-life-cycle-training-aritra-ghosh/> [accessed 2023-09-20]
31. Sachdeva PS, Barreto R, von VC, Kennedy CJ. Assessing annotator identity sensitivity via item response theory: a case study in a hate speech corpus. USA: Association for Computing Machinery; 2022 Presented at: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency; 2022 June 21-24; Seoul Republic of Korea. [doi: [10.1145/3531146.3533216](https://doi.org/10.1145/3531146.3533216)]

32. Chan A. GPT-3 and InstructGPT: technological dystopianism, utopianism, and “Contextual” perspectives in AI ethics and industry. *AI Ethics* 2023;3(1):53-64. [doi: [10.1007/s43681-022-00148-6](https://doi.org/10.1007/s43681-022-00148-6)]
33. Bender EM, Gebru T, McMillan-Major A, Shmitchell S. On the dangers of stochastic parrots: can language models be too big? USA: Association for Computing Machinery; 2021 Presented at: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency; 2021 March 3-10; Virtual Event, Canada p. 610-623. [doi: [10.1145/3442188.3445922](https://doi.org/10.1145/3442188.3445922)]
34. Ntoutsi E, Fafalios P, Gadiraju U, Iosifidis V, Nejdil W, Vidal M, et al. Bias in data-driven artificial intelligence systems—An introductory survey. *WIRES Data Min Knowl Discov* 2020;10(3):e1356. [doi: [10.1002/widm.1356](https://doi.org/10.1002/widm.1356)]
35. Sun T, Gaut A, Tang S, Huang Y, ElSherief M, Zhao J, et al. Mitigating gender bias in natural language processing: literature review. : Association for Computational Linguistics; 2019 Presented at: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics; 2019 July 28- August 2; Florence, Italy p. 1630-1640. [doi: [10.18653/v1/p19-1159](https://doi.org/10.18653/v1/p19-1159)]
36. Hovy D, Prabhumoye S. Five sources of bias in natural language processing. *Lang Linguist Compass* 2021;15(8):e12432 [FREE Full text] [doi: [10.1111/lnc3.12432](https://doi.org/10.1111/lnc3.12432)] [Medline: [35864931](https://pubmed.ncbi.nlm.nih.gov/35864931/)]
37. Steed R, Caliskan A. Image representations learned with unsupervised pre-training contain human-like biases. 2021 Presented at: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency; 2021 March 3-10; Virtual Event, Canada p. 701-713. [doi: [10.1145/3442188.3445932](https://doi.org/10.1145/3442188.3445932)]
38. How it feels to be sexually objectified by an AI. *MIT Technol Rev*. URL: <https://www.technologyreview.com/2022/12/13/1064810/how-it-feels-to-be-sexually-objectified-by-an-ai/> [accessed 2023-09-20]
39. Castaneda J, Jover A, Calvet L, Yanes S, Juan A, Sainz M. Dealing with gender bias issues in data-algorithmic processes: a social-statistical perspective. *Algorithms Multidisciplinary Digital Publishing Institute* 2022;15(9):303. [doi: [10.3390/a15090303](https://doi.org/10.3390/a15090303)]
40. Wellner G, Rothman T. Feminist AI: can we expect our AI systems to become feminist? *Philos Technol* 2020;33(2):191-205. [doi: [10.1007/s13347-019-00352-z](https://doi.org/10.1007/s13347-019-00352-z)]
41. Zhou J, Zhang Y, Luo Q, Parker A, De CM. Synthetic lies: understanding AI-generated misinformation and evaluating algorithmic and human solutions. : ACM; 2023 Presented at: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems; 2023 April 23 - 28; Hamburg, Germany p. 1-20. [doi: [10.1145/3544548.3581318](https://doi.org/10.1145/3544548.3581318)]
42. Vinay R, Spitale G, Biller-Andorno N, Germani F. Emotional manipulation through prompt engineering amplifies disinformation generation in AI large language models. *Computer Science > Artificial Intelligence* 2024:1-14. [doi: [10.48550/arXiv.2403.03550](https://doi.org/10.48550/arXiv.2403.03550)]
43. Four years later, AI language dataset created by brown graduate students goes viral. *Brown Univ*. 2023. URL: <https://www.brown.edu/news/2023-04-25/open-web-text> [accessed 2023-09-20]
44. Patenaude J, Legault G, Beauvais J, Bernier L, Béland J, Boissy P, et al. Framework for the Analysis of Nanotechnologies? Impacts and ethical acceptability: basis of an interdisciplinary approach to assessing novel technologies. *Sci Eng Ethics* 2025;21(2):293-315. [doi: [10.1007/s11948-014-9543-y](https://doi.org/10.1007/s11948-014-9543-y)]
45. Taebi B. Bridging the gap between social acceptance and ethical acceptability. *Risk Anal* 2017;37(10):1817-1827. [doi: [10.1111/risa.12734](https://doi.org/10.1111/risa.12734)] [Medline: [27862106](https://pubmed.ncbi.nlm.nih.gov/27862106/)]
46. Hacker P. A legal framework for AI training data from first principles to the artificial intelligence act. *Law Innov Technol Routledge* 2021;13(2):257-301. [doi: [10.1080/17579961.2021.1977219](https://doi.org/10.1080/17579961.2021.1977219)]
47. Artificial intelligence act: deal on comprehensive rules for trustworthy AI. *News | European Parliament*. 2023. URL: <https://www.europarl.europa.eu/news/en/press-room/20231206IPR15699/artificial-intelligence-act-deal-on-comprehensive-rules-for-trustworthy-ai> [accessed 2024-01-17]
48. Goldstein J, Sastry G, Musser M, DiResta R, Gentzel M, Sedova K. Generative language models and automated influence operations: emerging threats and potential mitigations. *arXiv*. 2023. URL: <http://arxiv.org/abs/2301.04246> [accessed 2023-09-20]
49. Lessons Learned on Language Model Safety and Misuse. URL: <https://openai.com/research/language-model-safety-and-misuse> [accessed 2023-09-20]
50. Ganguli D, Lovitt L, Kernion J, Askell A, Bai Y, Kadavath S, et al. Red teaming language models to reduce harms: methods, scaling behaviors, and lessons learned. *arXiv*. URL: <http://arxiv.org/abs/2209.07858> [accessed 2023-09-20]
51. The AI Detection Arms Race Is On | WIRED. URL: <https://www.wired.com/story/ai-detection-chat-gpt-college-students/> [accessed 2023-09-20]
52. Smith T. Arms race instability and war. *J Confl Resolut SAGE Publications Inc* 1980;24(2):253-284. [doi: [10.1177/002200278002400204](https://doi.org/10.1177/002200278002400204)]
53. ChatGPT and the fight against disinformation: how AI is changing the game. *Culturico*. URL: <https://culturico.com/2023/03/04/chatgpt-and-the-fight-against-disinformation-how-ai-is-changing-the-game/> [accessed 2023-09-20]
54. Germani F, Biller-Andorno N. How to counter the anti-vaccine rhetoric: filling information voids and building resilience. *Hum Vaccin Immunother* 2022;18(6):2095825 [FREE Full text] [doi: [10.1080/21645515.2022.2095825](https://doi.org/10.1080/21645515.2022.2095825)] [Medline: [35802046](https://pubmed.ncbi.nlm.nih.gov/35802046/)]
55. Convention on cybercrime ETS - No. 185. 2001. Council of Europe. 2001. URL: <https://rm.coe.int/1680081561> [accessed 2024-09-17]
56. How Digital Identity can Protect Against Misuse of AI. URL: <https://oneid.uk/news-and-events/how-digital-identity-can-protect-against-misuse-of-ai> [accessed 2023-09-20]

57. Ahmad W, Berg R, Kim S. Combating Fake News with Digital Identity Verification. URL: <https://groups.csail.mit.edu/mac/classes/6.805/student-papers/fall17-papers/FakeNews.pdf> [accessed 2024-09-17]
58. DeVerna MR, Yan HY, Yang KC, Menczer F. Artificial intelligence is ineffective and potentially harmful for fact checking. arXiv. 2023. URL: <http://arxiv.org/abs/2308.10800> [accessed 2023-09-20]
59. Sebastian G. Exploring ethical implications of ChatGPT and other AI Chatbots and regulation of disinformation propagation. SSRN 2023:1-16. [doi: [10.2139/ssrn.4461801](https://doi.org/10.2139/ssrn.4461801)]
60. About Community Notes on X | X Help. URL: <https://help.twitter.com/en/using-x/community-notes> [accessed 2023-09-21]
61. Directorate general for parliamentary research services. Regulating disinformation with artificial intelligence: effects of disinformation initiatives on freedom of expression and media pluralism. European Parliament. LU: Publications Office; 2019. URL: <https://data.europa.eu/doi/10.2861/003689> [accessed 2023-09-21]
62. Meyer T. Regulating Disinformation with Artificial Intelligence?. URL: [https://www.europarl.europa.eu/RegData/etudes/STUD/2019/624279/EPRS_STU\(2019\)624279_EN.pdf](https://www.europarl.europa.eu/RegData/etudes/STUD/2019/624279/EPRS_STU(2019)624279_EN.pdf) [accessed 2024-09-17]
63. Harris ED. Governance of Dual-Use Technologies. URL: <https://www.amacad.org/publication/governance-dual-use-technologies-theory-and-practice/section/3> [accessed 2024-09-17]
64. Appedu S, Hensley MK. Problematizing the role of information literacy in disinformation, dialogue, the healing of democracy. In: Sietz B, editor. *Inf Lit Time Transform*. Michigan: LOEX Press; 2021.
65. Ringing the Alarm Bell with Federico Germani. URL: <https://www.andybusam.com/ringing-the-alarm-bell-with-federico-germani/> [accessed 2023-09-21]
66. Redaelli S, Biller-Andorno N, Gloeckler S, Brown J, Spitale G, Germani F. Mastering critical thinking skills is strongly associated with the ability to recognize fakeness and misinformation. SocArXiv (OSF) 2024. [doi: [10.31235/osf.io/hsz6a](https://doi.org/10.31235/osf.io/hsz6a)]
67. Jones-Jang SM, Mortensen T, Liu J. Does media literacy help identification of fake news? Information literacy helps, but other literacies don't. *Am Behav Sci* 2019;65(2):371-388. [doi: [10.1177/0002764219869406](https://doi.org/10.1177/0002764219869406)]
68. De PS, Heravi B. Information literacy and fake news: how the field of librarianship can help combat the epidemic of fake news. *J Acad Librariansh* 2020;46(5):102218. [doi: [10.1016/j.acalib.2020.102218](https://doi.org/10.1016/j.acalib.2020.102218)]
69. Willingham DT. Ask the cognitive scientist: how can educators teach critical thinking? *Am Educ American Federation of Teachers, AFL-CIO* 2020;3(41):44.
70. Gaillard S, Oláh ZA, Venmans S, Burke M. Countering the cognitive, linguistic, and psychological underpinnings behind susceptibility to fake news: a review of current literature with special focus on the role of age and digital literacy. *Front Commun*. 2021. URL: <https://www.frontiersin.org/articles/10.3389/fcomm.2021.661801> [accessed 2023-09-26]
71. Allner IB. Teaching of Information Literacy: Collaboration Between Teaching Faculty and Librarians. US: BiblioBazaar; 2011.
72. Johnston B, Webber S. Information literacy in higher education: a review and case study. *Stud High Educ Routledge* 2003;28(3):335-352. [doi: [10.1080/03075070309295](https://doi.org/10.1080/03075070309295)]
73. Burclaff N, Johnson C. Teaching Information Literacy via Social Media: An Exploration of Connectivism. URL: https://www.researchgate.net/publication/316187027_Teaching_Information_Literacy_via_Social_Media_An_Exploration_of_Connectivism [accessed 2024-09-17]
74. Fake news created by artificial intelligence is difficult to recognize. They seem more credible to Internet users than messages created by humans. *Bizness*. URL: <http://biznes.newseria.pl/news/fake-newsy-stworzone-przez.p919781558> [accessed 2023-09-22]
75. Tiernan P, Costello E, Donlon E, Parysz M, Scriney M. Information and media literacy in the age of AI: options for the future. *Educ Sci Multidisciplinary Digital Publishing Institute* 2023;13(9):906. [doi: [10.3390/educsci13090906](https://doi.org/10.3390/educsci13090906)]
76. Hwang Y, Ryu JY, Jeong S. Effects of disinformation using deepfake: the protective effect of media literacy education. *Cyberpsychol Behav Soc Netw* 2021;24(3):188-193. [doi: [10.1089/cyber.2020.0174](https://doi.org/10.1089/cyber.2020.0174)] [Medline: [33646021](https://pubmed.ncbi.nlm.nih.gov/33646021/)]
77. WHO Kicks off Deliberations on Ethical Framework and Tools for Social Listening and Infodemic Management. URL: <https://www.who.int/news/item/10-02-2023-who-kicks-off-deliberations-on-ethical-framework-and-tools-for-social-listening-and-infodemic-management> [accessed 2023-09-22]
78. Pause giant AI experiments: an open letter. Future Life Inst. URL: <https://futureoflife.org/open-letter/pause-giant-ai-experiments/> [accessed 2023-10-02]

Abbreviations

AI: artificial intelligence

Edited by K El Emam, B Malin, A Blasimme; submitted 09.10.23; peer-reviewed by E Pertwee, S Gordon, E Wilhelm; comments to author 13.01.24; revised version received 22.01.24; accepted 28.07.24; published 15.10.24.

Please cite as:

Germani F, Spitale G, Biller-Andorno N

The Dual Nature of AI in Information Dissemination: Ethical Considerations

JMIR AI 2024;3:e53505

URL: <https://ai.jmir.org/2024/1/e53505>

doi: [10.2196/53505](https://doi.org/10.2196/53505)

PMID: [39405099](https://pubmed.ncbi.nlm.nih.gov/39405099/)

©Federico Germani, Giovanni Spitale, Nikola Biller-Andorno. Originally published in JMIR AI (<https://ai.jmir.org>), 15.10.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR AI, is properly cited. The complete bibliographic information, a link to the original publication on <https://www.ai.jmir.org/>, as well as this copyright and license information must be included.

Original Paper

Privacy-Preserving Federated Survival Support Vector Machines for Cross-Institutional Time-To-Event Analysis: Algorithm Development and Validation

Julian Späth¹, MSc; Zeno Sewald², BSc; Niklas Probul¹, MSc; Magali Berland³, PhD; Mathieu Almeida³, PhD; Nicolas Pons³, PhD; Emmanuelle Le Chatelier³, PhD; Pere Ginès^{4,5,6,7}, MD, PhD; Cristina Solé^{4,5,6}, MD; Adrià Juanola^{4,5,6}, MD, PhD; Josch Pauling², PhD; Jan Baumbach¹, Prof Dr

¹Institute for Computational Systems Biology, University of Hamburg, Hamburg, Germany

²LipiTUM, Chair of Experimental Bioinformatics, TUM School of Life Sciences, Technical University of Munich, Freising, Germany

³MetaGenoPolis, INRAE, Université Paris-Saclay, Jouy-en-Josas, France

⁴Liver Unit, Hospital Clínic de Barcelona, Barcelona, Spain

⁵Institut d'Investigacions Biomèdiques August Pi i Sunyer (IDIBAPS), Barcelona, Spain

⁶Centro de Investigación en Red de Enfermedades hepáticas y Digestivas (CIBERehD), Madrid, Spain

⁷Faculty of Medicine and Health Sciences, University of Barcelona, Barcelona, Spain

Corresponding Author:

Julian Späth, MSc

Institute for Computational Systems Biology

University of Hamburg

Notkestrasse 9

Hamburg, 22607

Germany

Phone: 49 15750665331

Email: julian.alexander.spaeth@uni-hamburg.de

Abstract

Background: Central collection of distributed medical patient data is problematic due to strict privacy regulations. Especially in clinical environments, such as clinical time-to-event studies, large sample sizes are critical but usually not available at a single institution. It has been shown recently that federated learning, combined with privacy-enhancing technologies, is an excellent and privacy-preserving alternative to data sharing.

Objective: This study aims to develop and validate a privacy-preserving, federated survival support vector machine (SVM) and make it accessible for researchers to perform cross-institutional time-to-event analyses.

Methods: We extended the survival SVM algorithm to be applicable in federated environments. We further implemented it as a FeatureCloud app, enabling it to run in the federated infrastructure provided by the FeatureCloud platform. Finally, we evaluated our algorithm on 3 benchmark data sets, a large sample size synthetic data set, and a real-world microbiome data set and compared the results to the corresponding central method.

Results: Our federated survival SVM produces highly similar results to the centralized model on all data sets. The maximal difference between the model weights of the central model and the federated model was only 0.001, and the mean difference over all data sets was 0.0002. We further show that by including more data in the analysis through federated learning, predictions are more accurate even in the presence of site-dependent batch effects.

Conclusions: The federated survival SVM extends the palette of federated time-to-event analysis methods by a robust machine learning approach. To our knowledge, the implemented FeatureCloud app is the first publicly available implementation of a federated survival SVM, is freely accessible for all kinds of researchers, and can be directly used within the FeatureCloud platform.

(JMIR AI 2024;3:e47652) doi:[10.2196/47652](https://doi.org/10.2196/47652)

KEYWORDS

federated learning; survival analysis; support vector machine; machine learning; federated; algorithm; survival; FeatureCloud; predict; predictive; prediction; predictions; Implementation science; Implementation; centralized model; privacy regulation

Introduction

Accessing data to apply machine learning (ML) in biomedical settings is still challenging [1]. Large amounts of data exist in clinical settings but are scattered across numerous institutions. Due to strict privacy regulations, such as the General Data Protection Regulation (GDPR), this data cannot be easily shared or collected at a central institution [2]. This causes hurdles for cross-institutional biomedical analyses that depend on highly sensitive patient data. One example is time-to-event analysis, aiming to find parameters that prolong or shorten the time until a particular event, such as death, occurs [3]. In these studies, the event of interest does not necessarily occur for all samples, increasing the need for large sample sizes [4]. Until today, the need for large sample sizes and heterogeneous data for time-to-event studies is still mainly solved through traditional data sharing, leading to the central collection of various deidentified and anonymized data sets from different centers. Since using anonymized data in the training of ML models tends to weaken model performance [5], this comes with a tradeoff of data privacy and data quality, accelerating the need for alternative methods that keep data private and ensure the quality of the data [6].

In recent years, federated learning (FL) has become a feasible alternative to central data collection by enabling the training of models on distributed data sets. Instead of sharing sensitive data with a central institution, in FL, only insensitive model parameters are shared with a central aggregation server [7,8]. Therefore, each participating party calculates its own model with local model parameters on their local data. These local model parameters are then shared with the aggregator and aggregated into a global model. Afterward, the global model is shared again with each participant and can be updated in another iteration. The first and probably most widely used aggregation approach is the federated average [9], calculating the weighted mean of the exchanged model parameters. Besides using different aggregation approaches, FL can also be distinguished between horizontal and vertical learning, as well as cross-device and cross-silo learning. Horizontal learning describes FL on data with the same features but different samples, while vertical learning performs on the same samples but with different features between the participating parties. Cross-device FL trains models across millions of participants (such as mobile phones), cross-silo FL, on the other hand, focuses on a few clients only, such as hospitals or research institutes [10].

Especially in combination with privacy-enhancing techniques (PETs), model parameters can be exchanged securely, such that a global aggregator or potential attacker cannot even see the local parameters of each participant [11]. This secure exchange of model parameters is necessary to comply with the GDPR, as even local models can be considered personal data [12]. Therefore, FL enables the training on a significantly larger data set compared with single-institution scenarios. While federated algorithms still often struggle with communication efficiency,

the significantly increased amount of data can offset this performance issue, making FL a serious competitor to classical ML. Additionally, since FL models are trained on a larger variety of data, they typically generalize better than traditional ML models and even generalize faster in some cases [13,14]. Many FL approaches are already published for biomedical applications, such as medical imaging analysis, genome-wide association studies, or gene expression analysis [15-17].

In addition to federated ML approaches, several federated time-to-event analysis algorithms have been introduced recently and confirmed their high potential for privacy-preserving analyses [18-21]. However, existing approaches solely cover traditional statistical methods such as the estimation of survival functions and the Cox proportional hazards model. Modern ML algorithms for survival analysis, such as survival Support Vector Machines (SVMs), are not yet available in a federated fashion, even though SVMs belong to one of the most popular ML methods. If algorithms are not available in federated scenarios, this might be a reason why researchers chose not to perform FL, if their favorite algorithms are not available. Many well-performing centralized algorithms are challenging to translate to a federated scenario while keeping sensitive data private. Another limitation of FL is communication efficiency. FL algorithms need to exchange the intermediate statistics with a central aggregator, which is especially inefficient for algorithms with many iterations. This inefficiency even increases when adding secure aggregation schemes, such as additive secret sharing. This PET ensures that only masked and encrypted model parameters are shared with the aggregating party, securing the local models from data leakage [18].

To address the lack of availability of federated time-to-event methods, we propose a privacy-preserving, horizontally federated, cross-silo survival SVM based on the survival analysis package *scikit-survival* [22]. Compared with other existing time-to-event methods, such as the Cox proportional hazard model, the survival SVM allows an actual prediction of the time until an event happens. It can be used to predict the risk of individual samples, which is not possible in univariate time-to-event algorithms and is not the aim of the Cox proportional hazards model. Therefore, to the best of our knowledge, it is the first freely available federated survival prediction method. We implemented the algorithm as an app in the FeatureCloud platform to make it publicly accessible and to minimize the hurdles of FL infrastructure [23]. Based on a combination of FL and additive secret sharing, we show on 3 benchmark data sets, that our approach achieves highly similar results compared with central data analysis. Additionally, we apply it to a set of real-world microbiome data sets to demonstrate its applicability to original clinical data.

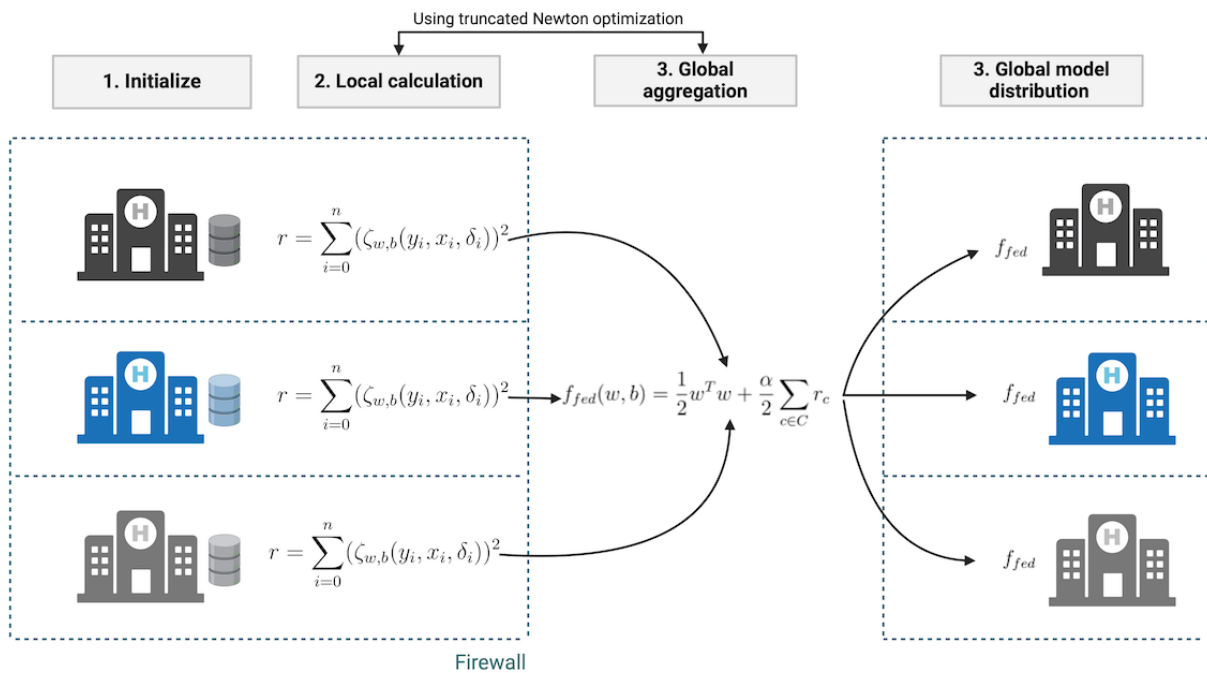
Methods

Here, we propose the adapted algorithm for the federated survival SVM, describe its implementation as a FeatureCloud app, and explain how we evaluated its performance.

Federated Survival SVM

We extended the regression objective of *scikit-survival*'s FastSurvivalSVM without ranking to be applicable in federated environments [24]. As shown in Figure 1, instead of calculating the sum of the squared ζ -function centrally, it is calculated at each site, with the feature vector x_i , the survival time $y_i > 0$, and the binary event indicator δ_i . Each site's local sum of squared ζ -function is then sent to a global aggregator and summed up to the global sum of squared ζ -function. The below equations show the central objective function and our corresponding federated objective function, with C being the set of all participating clients.

Figure 1. Federated calculation of a survival support vector machine (SVM). Each site calculates the sum of squares locally and sends it to the global aggregation server. The aggregation server aggregates the local sum of squares by summing them up to the global sum of squares. The objective function is minimized in a federated fashion by a truncated Newton approach. After convergence, the global model is distributed to all participating clients.



FeatureCloud

We developed an FL app on the FeatureCloud platform to make our approach publicly available. To develop this app, we used the app template and application programming interface provided by FeatureCloud [25]. Using the *scikit-survival* package and Python, we implemented our algorithm, put it into the FeatureCloud app template, and published it in the FeatureCloud artificial intelligence store. It can be used with other apps in a workflow or standalone using the platform. Our code is entirely open source.

In FeatureCloud, 1 participating client also takes the aggregating role and is called the coordinator. The app is implemented as a state machine, meaning that the app switches between states to

Mathematically, our federated formula leads to the same solution as the centralized calculation of the objective function. Similar to the centralized analysis, a truncated Newton method (such as Newton-CG) can be used to optimize the objective function. For this, in each iteration, the gradient and Hessian matrix of each client are also sent to the global aggregator to sum them up to the global gradient and Hessian matrix. To reduce potential privacy leakage from the exchanged data, the implementation of the federated algorithm should support a secure aggregation scheme that hides the locally exchanged data from attackers or the global aggregation server.

perform different tasks. All states and their transitions are shown in Multimedia Appendix 1. After reading the local data and config files, minimizing the objective function using a federated Newton conjugate gradient is performed iteratively. Therefore, the local gradient and Hessian matrices are calculated and sent to the coordinator. The coordinator aggregates these data to obtain the global matrices, updates the weight vector ω , and broadcasts it to all clients. This is repeated until convergence.

A considerable advantage of the FeatureCloud platform is its native support of 2 very popular PETs, such as secure multiparty computation (SMPC). For applying SMPC, FeatureCloud supports a secure aggregation scheme for hiding locally exchanged parameters using additive secret sharing [26]. Through this, the exchanged local models are protected, and

only the global aggregations are visible to attackers, clients, and the global aggregator. This is achieved by splitting the value that needs to be exchanged with the global aggregator into n shards, where n is the number of participating clients, and the sum of these n shards would result in the actual value [23]. Each shard is encrypted using a public key of each participant. These encrypted shards are shared with the global aggregator, sending them to the corresponding client holding the private key. The clients decrypt the received shards, sum them up, and send them back to the global aggregator, which sums up all received sums. This final sum results in the actual, nonhidden, global aggregate.

Ethical Considerations

According to German regulations, for our retrospective study performed on publicly available data or data with explicit consent, approval from an ethical committee was not required.

Evaluation

We evaluated our approach using the developed FeatureCloud app on 3 benchmark data sets, all available via the *scikit-survival* package. The breast cancer data set (BRCA) [27] contains the gene expression profiling of microarray experiments from 198 primary breast tumors, originally used to validate a 76-gene prognostic signature able to predict distant metastases in lymph node-negative patients with breast cancer. The German Breast Cancer Study Group 2 data set (GBSG2) [28] contains data from a multicenter randomized clinical trial to compare the effectiveness of 3 versus 6 cycles of cyclophosphamide, methotrexate, and fluorouracil on recurrence-free and overall survival of 686 women. The observed parameters were hormonal therapy (yes or no), age of the patients, menopausal status (pre vs post), tumor size (in mm), tumor grade, number of positive

tumor nodes, progesterone receptor (in fmol), and estrogen, as well as the censoring indicator and recurrence-free survival time (in days). The Worcester Heart Attack Study data set (WHAS500) [29] contains data from 500 patients with acute myocardial infarction, collected during thirteen 1-year periods. Parameters were age, gender, initial heart rate, initial systolic and diastolic blood pressure, body mass index, history of cardiovascular disease, atrial fibrillation, cardiogenic shock, congestive heart complications, complete heart block, myocardial infarction order and type, vital status, and total length of follow-up.

Additionally, we evaluated our algorithm on a recent, high-dimensional gut microbiome data set from the Hospital Clinic of Barcelona, containing data from 150 patients with liver cirrhosis [30]. The data set was aimed at assessing the predicting role of the gut microbiome for the survival of the patients in the context of liver cirrhosis, using shotgun metagenomic sequencing performed on fecal DNA isolated from stool samples. A former version of the data has been previously analyzed with a different methodology [30]. For this study, the Metagenomic Species Pangenome (MSP) was used to identify and quantify microbial species associated with the IGC2 reference catalog [31]. MSPs are clusters of coabundant genes (minimum size >100 genes) used as a proxy for microbial species, reconstructed from 1601 metagenomes to 1990 MSP species [32]. MSP abundances were estimated as the mean abundance of their 100 marker genes, as far as at least 20% of these genes are detected. The MSP abundance table was then normalized in each sample by dividing its abundance by the sum of MSP abundances detected in the sample. Further details regarding the data sets are shown in Table 1.

Table 1. Overview of all data sets. Our 4 evaluation data sets differ greatly in the number of samples, features, events, and censored individuals. Features indicate the number of clinical variables or microbial species abundance in the data set; median follow-up indicates the median follow-up time of the patients in days; events indicate the number of patients for whom the event of interest was observed during observation time; and censored indicates the number of patients for whom the event of interest was not observed during observation time.

Data set	Samples, n	Features, n	Median follow-up (days)	Events, n	Censored, n	End point
BRCA	198	84	4384.0	51	147	Presence of metastases
GBSG2	686	11	1084.0	299	387	Recurrence-free survival
WHAS500	500	16	631.5	215	285	Death
Microbiome	150	1995	416.0	51	99	Death

^aBRCA: breast cancer data set.

^bGBSG2: German Breast Cancer Study Group 2 data set.

^cWHAS500: Worcester Heart Attack Study data set.

We one-hot encoded nonbinary categorical features. For each data set, we created either 1 client (100%) as the centralized scenario, 3 clients (20%, 50%, and 30%) as the multicentric imbalanced scenario, and 5 clients (20% each) as the multicentric balanced scenario, and we split the data accordingly.

To evaluate the accuracy of our model, we used the Harrell concordance index, which was developed as a generalization of the area under the receiver operating characteristic curve for

time-to-event models [33]. It corresponds to the probability of concordance between observed and predicted survival based on each pair of individuals. A c-index of 0.5 means that the model performs as well as a random guess, and a c-index of 1.0 means that the model predicts perfectly well.

After preprocessing, we performed a 3 × 3-fold cross-validation (CV) for a FeatureCloud workflow consisting of a federated normalization, the federated survival SVM, and a federated survival evaluation (c-index). We then compared our results

with the centralized analysis of every client and the merged data set (simulating a central data collection). Centralized analysis was performed using *scikit-survival*'s FastSurvivalSVM with a rank ratio of 0, α of 0.0001, true fit intercept, and a maximum of 50 iterations. The same hyperparameters were used for the federated analysis, respectively.

Privacy

FeatureCloud supports several properties to increase the privacy and security of the computations. One important step is that FL projects can be only executed with invited participants. For this, a unique and secret code is needed to join the project. Every participant can see the workflow and each individually executed FeatureCloud app that will run in the workflow. As FeatureCloud apps are open source, even the executed code of the apps can be examined.

The execution of apps and workflows in FeatureCloud is containerized and strictly monitored. Due to the containerization, individual apps are not allowed to establish a connection to the internet, which prevents the extraction of data from malicious code. Even though the communication between clients does not contain sensitive patient information, it is RSA (Rivest–Shamir–Adleman) encrypted through the standard HTTPS protocol. This prevents unauthorized third parties from gaining insights into parameters exchanged during training.

Exchanged parameters from each individual site are masked through the secure aggregation scheme, hiding the intermediate statistics from other participating clients and the global aggregator. This efficiently addresses the problem of local models considered as personal data in GDPR [18].

Our federated survival SVM app currently uses a hybrid approach of SMPC and FL. This hybrid approach increases the privacy of the exchanged local parameters from both participants and potential attackers, as explained in the methods section.

Differential privacy (DP) [34] is not yet supported by FeatureCloud but is currently in development and could be added to the algorithm as an additional layer to improve privacy. However, as the app trains a linear model, it is less prone to overfit, reducing the surface for potential membership and attribute inference attacks [35]. In DP, noise is added to the model parameters during the training process to guarantee a mathematically quantifiable amount of privacy for each sample. While this comes with large advantages regarding privacy, the application of DP has also various weaknesses. The addition of noise lowers the performance of the model significantly, especially when applying the amount of noise necessary for a meaningful level of privacy [36]. Further, this guarantee only is applicable for a limited number of interactions with the

resulting model. As the final model is distributed to all participants, they can interact with the model arbitrarily, making the privacy guarantee void, thus not warranting an inclusion in this analysis.

A PET not supported by FeatureCloud currently is homomorphic encryption (HE), which allows the computation of the model on encrypted values, making sharing of data even more secure. While this is great in theory, it actually gains very little benefit in this analysis scenario. The data we share is already nonsensitive and through the use of SMPC, we can hide not only the data but the data's origin. This is why FeatureCloud currently supports SMPC instead of HE.

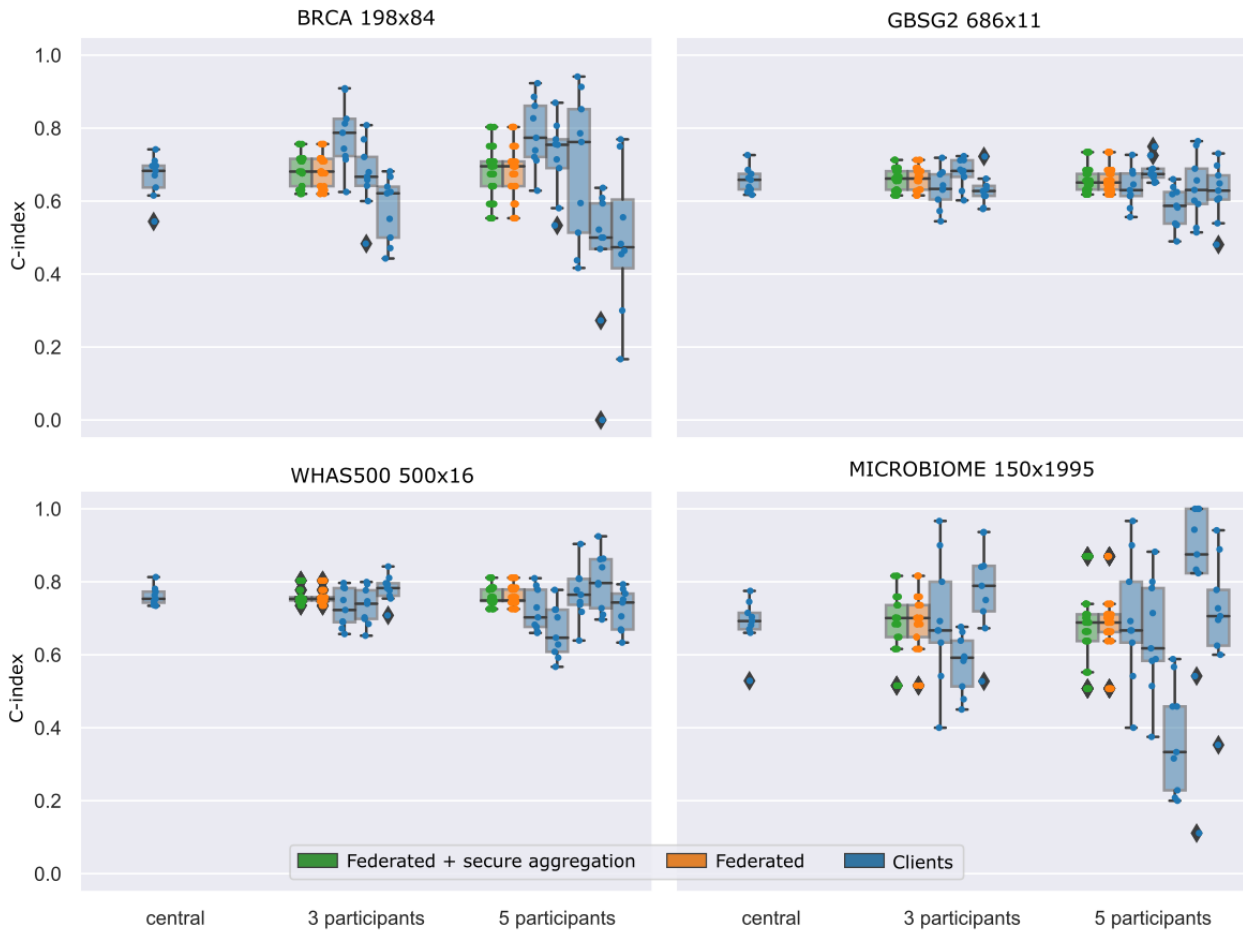
Our implementation of the federated survival SVM app uses all the functionalities offered by FeatureCloud and does not deviate from these best practices.

Results

Performance

Our workflow delivered a highly similar model performance and model parameters for all federated analyses compared with the ones performed on the corresponding centralized data sets. The resulting *c*-indices to estimate the performance of our time-to-event models are depicted in Figure 2 [33]. For each data set (subplot), we show a boxplot consisting of the evaluated *c*-index for each CV split of our federated workflow with secure aggregation (green), federated workflow without secure aggregation (orange), and centralized calculation for each individual client (blue). The CV results show that our federated as well as the federated and secure aggregation approach perform highly similar to the centralized estimates. The calculation of the federated *c*-index in FeatureCloud causes small deviances in the *c*-index between centralized and federated. This is because FeatureCloud calculates a local *c*-index and aggregates to the mean *c*-indices of all sites. Therefore, it does not lead to the same *c*-index as a central computation would. The mean *c*-indices for the 4 data sets are in the range between 0.658 (GDBG2) and 0.76 (WHAS500). In contrast to the accuracy, achieving very high *c*-indices is rather difficult and depends very much on the problem. In a bioinformatics context, the lowest *c*-index of 0.658 (GDBG2) can be considered as moderate. The model achieves discrimination between individuals with different survival outcomes. However, it might not be of clinical utility and needs further refinement. The *c*-index of 0.76 (WHAS500) on the other hand, can be considered as good and has predictive value. Improving the predictive value of the models and increasing *c*-index was out of the scope of this work. A complete table of the results is available in [Multimedia Appendix 2](#).

Figure 2. Comparison of federated and centralized analysis. The boxplots show the evaluated c-indices (3×3 -fold cross validation) of the central, 3 participants, and 5 participants analysis (rows). For each scenario, we compared the federated and secure aggregation approach (green), the federated-only approach (orange), and the performance of every single site (blue). BRCA: breast cancer data set; GBSG2: German Breast Cancer Study Group 2 data set; WHAS500: Worcester Heart Attack Study data set.



The model weights are nearly identical, with a maximum difference of only 0.001 and a mean difference of 0.0002 (Multimedia Appendices 1 and 3). These tiny differences between the weights of the central model and our model are negligible, as they do not change the overall prediction results and still lead to equal c-indices. The resulting model is therefore almost identical to the one that was trained on central data. A useful property of the linear survival SVM is, that the model weights can be used as a feature importance measure, which is also supported in our approach.

Besides calculating the feature importance from model weights directly, our federated survival SVM app uses Shapley additive explanations (SHAP), an explainable artificial intelligence framework for the interpretation of ML models [37]. Using SHAP, we compared the final models of the central, federated without secure aggregation, and federated with secure aggregation runs. For each data set, the SHAP shows highly similar model interpretations with a mean Pearson correlation of 0.991 between the central and the federated model without secure aggregation, and a mean Pearson correlation of 0.985 between the central model and the federated model with secure aggregation. A slightly worse correlation in the secure aggregation model is expected, as the masking of local parameters leads to floating-point issues. The worst correlation

is shown in the microbiome data set (0.964), which can be explained by the high correlation between features in this data set. The results of the SHAP correlation analysis are listed in Multimedia Appendix 4 and the corresponding SHAP beeswarm plots are available in Multimedia Appendix 5.

Our results further demonstrate the importance of large data sets, as the performance of the locally trained models on single clients (smaller sample size) shows a much higher variance than our federated models. If 5 institutes combine their small data sets, they can perform a much more reliable time-to-event analysis compared with isolated institutions. This further supports the high practical value of FL in real-world clinical time-to-event analysis, especially for institutions with small sample sizes, homogenous cohorts, or only a few patients with rare diseases.

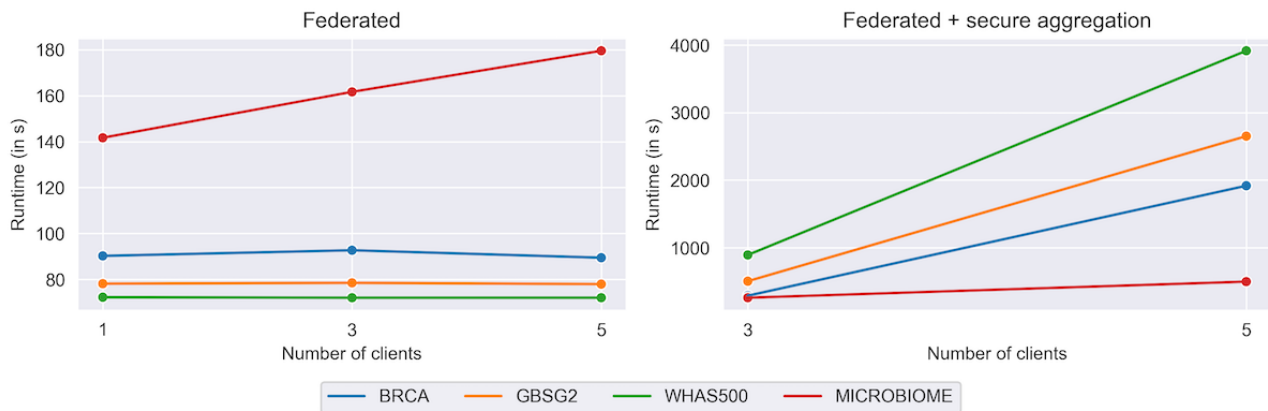
Runtime

As shown in Figure 3, the runtime largely depends on the data set. In the case of FL, the number of iterations and, therefore, the number of data exchanges are the bottleneck. While the federated-only approach has linear runtime, the runtime of federated and secure aggregation is much worse and increases with an increasing number of clients. As described in the FeatureCloud publication, providing better privacy by hiding

the exchanged parameters from the global aggregator, the simple additive secret sharing grows quadratic with the number of participants. Especially when many iterations and data

exchanges are needed, this has a bad influence on the runtime of the FL implementation.

Figure 3. Runtime analysis. The lines represent the runtime for each data set and the number of participating clients. The federated-only approach is depicted on the left, and the federated and secure aggregation approach is depicted on the right.



All results of the runtime analysis are shown in [Multimedia Appendix 6](#). Additionally, we performed the runtime analysis on a data set with a large sample size. As real-world time-to-event data sets are difficult to find, we used a synthetically generated, published data set from an example colon data set with 15,564 samples [38]. Our results show that our method scales well for large sample sizes, as the number of iterations is the bottleneck in FL ([Multimedia Appendix 7](#)).

FeatureCloud App

The app we developed can easily be used within the FeatureCloud platform. For this, a project coordinator creates a project, selects the app, and invites collaborators. Each participant installs FeatureCloud and joins the project. The app expects 2 CSV files as input, one for the training data and another for the test data. A config file can be used to define hyperparameters and other descriptors, such as the time and event label columns. After the federated computation has finished, each client receives the globally trained model as a pickle file, as well as a prediction file containing all predictions on the local test data set. The app can also be used in a FeatureCloud workflow, supporting various preprocessing methods, such as CV, normalization, feature selection, one-hot encoding, and subsequent evaluation of survival models using the c-index.

The requirements for running the survival SVM app are the same as for executing the FeatureCloud platform. It requires a stable internet connection to exchange the incentive model parameters with the central aggregator and to run the app on the website. Docker needs to be installed on a Mac, Linux, or Windows computer with the corresponding requirements for running Docker [39]. Moreover, enough memory should be available to process the data set. This depends mainly on the data set size, and not on the algorithm itself.

Discussion

Principal Findings

Our federated survival SVM has been demonstrated to offer a highly viable alternative to centralized data collection in a time-to-event analysis. It achieves comparable levels of accuracy without compromising the privacy of highly sensitive patient data. This makes it a compelling solution for organizations seeking to safeguard sensitive data while still gaining the benefits of advanced analysis and the application of ML. Through its availability as a FeatureCloud app, the platform takes care of deployment and federated infrastructures, making it directly usable with little programming knowledge. The results of the real-world microbiome data set are promising and show that FL might be an accelerator in microbiome research and the analysis of time-to-event microbiome data sets. Using FL combined with additive secret sharing, our approach can be currently considered GDPR compliant and, therefore, practically usable in real clinical time-to-event studies [12].

Comparison to Existing Work

Only a few federated survival analysis approaches were developed in recent years, such as the distributed Cox proportional hazards model WebDISCO or an approach for federated survival curves using multiparty HE [18,20]. In a recent study about privacy-aware multi-institutional time-to-event analysis, it was criticized that the existing work was mainly focusing on theoretical solutions, rather than practical [21]. Therefore, lack of usability was a huge issue that was addressed by the authors, who developed the platform “Partea” [21]. The platform supports the Kaplan-Meier estimator for survival curve estimation [40], Nelson-Aalen estimator for cumulative hazard ratios [41], and Cox proportional hazards model for survival regression [42]. Compared with “Partea,” FeatureCloud does not only address the execution of FL algorithms, but also development. The FeatureCloud developer application programming interface for implementing FL algorithms that can be executed through FeatureCloud and published in the App Store is a huge advantage in terms of

development speed and also accessibility for the potential user group.

To our knowledge, the survival SVM FeatureCloud app is one of the first time-to-event analysis ML models implemented as a FL algorithm. This makes the accuracy (or c-index in our case) between the algorithms not directly comparable. However, similar to the existing solutions [20,21], our approach achieves almost identical results compared with the central algorithms.

Regarding runtime, univariate methods without iterations, such as Kaplan-Meier estimator, Nelson-Aalen estimator, or log-rank test are much more efficient in FL settings. However, these approaches cannot be used to analyze high dimensional data and multivariate settings. The efficiency of our approach is comparable to the iteratively trained Cox proportional hazard model, which is trained iteratively and requires communication and aggregation for every parameter update step.

Limitations

Our current approach does not support the more efficient ranking objective, as federated ranking is not trivial to implement. Instead, it is based on *scikit-survival*'s regression objective. Moreover, it solely supports the linear SVM and does not support the kernel SVM yet. Calculating a kernel matrix in a federated setting is not trivial, as it represents pairwise similarities (or distances) between the training data points. For supporting more complex, nonlinear relationships, this should be further investigated in the future. We still decided to implement and use a survival SVM in this work, as SVMs are very popular in health care and the only available time-to-event analysis ML model in *scikit-survival* that is not based on an ensemble approach. Ensemble models, such as random survival forests [43] or survival gradient boost, are both based on a set of survival trees. While ensemble models are also popular in time-to-event analysis, the federated aggregation of the local forests produces slightly worse results than centrally trained models in imbalanced scenarios [44]. A federated aggregation of each local tree, on the other hand, is computationally costly. The SVM in our implementation produces highly accurate results compared with central learning for model weights, c-index, and feature importance and can therefore lower the

burden of applying FL in health care (eg, microbiome analysis), as the participants can be sure that the results are equal to the ones they would obtain in a central setting.

FeatureCloud currently only supports a simple additive secret-sharing scheme, increasing runtime for calculations with many clients and iterations. This could be solved in the future by using a more efficient secret-sharing scheme, such as Shamir secret sharing, that is currently not supported by FeatureCloud [45]. By using FeatureCloud as the execution platform, our approach does not solve the still existing open problems of FL, such as fairness, debugging, and communication efficiency (especially when using secret sharing) [46]. Furthermore, there are attacks on FL architectures that cannot be prevented through the existing methods, such as privacy inference from the global model, and model or data poisoning [47]. It is therefore recommended to use the algorithms and FeatureCloud platform only with trusted parties.

Another limitation that comes from the FeatureCloud platform is data standardization. Data formatting and standards need to be discussed and determined in advance by the participants of the federated analysis. However, FeatureCloud provides the possibility to include federated data preprocessing applications in the workflow. While this does not remove the need for external communication of data standards, such as included features and naming conventions, it makes it straightforward to guarantee the same format and preprocessing for the used data before the actual model training process. Possible applications include imputation, normalization, train or test splitting, and CV [48,49].

Conclusions

In conclusion, we developed an open-source federated survival SVM that performs time-to-event analysis on geographically distributed data sets without sharing sensitive raw data. It is freely available in the FeatureCloud App Store. The trained models are almost identical compared with centrally trained survival SVMs. This extends the palette of existing federated time-to-event analysis approaches by another algorithm that can be applied to various problems.

Acknowledgments

This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement 826078. This publication reflects only the author's view and the European Commission is not responsible for any use that may be made of the information it contains (JB). This work was developed as part of the FeMAI project and is funded by the German Federal Ministry of Education and Research (BMBF) under grant 01IS21079 (NP) and by the Agence Nationale de la Recherche (ANR) under grant ANR-21-FAI1-0010. MB and MA were also supported by the grant ANR-11-DPBS-0001. JB was partially funded by his VILLUM Young Investigator Grant (13154). PG has received funds from the Instituto de Salud Carlos III through the Plan Estatal de Investigación Científica y Técnica y de Innovación, project references PI 16/00043 and PI 20/00579. These grants were cofunded by the European Regional Development Fund (FEDER) and also funded in part by an EU Horizon 20/20 Programme (H2020-SC1-2016-RTD), LIVERHOPE (731875). JKP is funded by the Bavarian State Ministry of Education and the Arts in the framework of the Bavarian Research Institute for Digital Transformation (bidt, grant LipiTUM)

Data Availability

The data sets generated and analyzed during this study are available in the GitHub repository [50]. The code for the implementation of the federated survival SVM is available in the GitHub repository [51]. The microbiome data set is not publicly available due to privacy regulations but is available from the corresponding author on reasonable request.

Conflicts of Interest

CS has received speaking fees from Abbvie and Grifols. PG has received research funding from Gilead & Grifols. PG has consulted or attended advisory boards for Gilead, RallyBio, SeaBeLife, Merck, Sharp and Dohme (MSD), Ocelot Bio, Behring, Roche Diagnostics International and Boehringer Ingelheim, and received speaking fees from Pfizer.

Multimedia Appendix 1

State workflow of the survival support vector machine (SVM) FeatureCloud app and difference between coefficients.
[\[DOCX File, 244 KB - ai_v3i1e47652_app1.docx\]](#)

Multimedia Appendix 2

C-indices of central, federated, and federated + secure aggregation analyses.
[\[XLSX File \(Microsoft Excel File\), 32 KB - ai_v3i1e47652_app2.xlsx\]](#)

Multimedia Appendix 3

Coefficients of the trained survival support vector machines (SVMs).
[\[XLSX File \(Microsoft Excel File\), 243 KB - ai_v3i1e47652_app3.xlsx\]](#)

Multimedia Appendix 4

Correlation of Shapley additive explanations (SHAP) values between central, federated, and federated + secure aggregation model.
[\[XLSX File \(Microsoft Excel File\), 10 KB - ai_v3i1e47652_app4.xlsx\]](#)

Multimedia Appendix 5

Shapley additive explanations (SHAP) beeswarm plots for the different models.
[\[ZIP File \(Zip Archive\), 25020 KB - ai_v3i1e47652_app5.zip\]](#)

Multimedia Appendix 6

Runtime of the federated survival support vector machine (SVM) training with 1, 3, and 5 clients.
[\[XLSX File \(Microsoft Excel File\), 11 KB - ai_v3i1e47652_app6.xlsx\]](#)

Multimedia Appendix 7

Runtime of the federated survival support vector machine (SVM) with 1, 3, and 5 clients of a large sample size synthetic data set.
[\[XLSX File \(Microsoft Excel File\), 10 KB - ai_v3i1e47652_app7.xlsx\]](#)

References

1. Adibuzzaman M, DeLaurentis P, Hill J, Benneyworth BD. Big data in healthcare—the promises, challenges and opportunities from a research perspective: a case study with a model database. *AMIA Annu Symp Proc* 2017;2017:384-392. [Medline: [29854102](#)]
2. Vlahou A, Hallinan D, Apweiler R, Argiles A, Beige J, Benigni A, et al. Data sharing under the general data protection regulation: time to harmonize law and research ethics? *Hypertension* 2021;77(4):1029-1035 [FREE Full text] [doi: [10.1161/HYPERTENSIONAHA.120.16340](#)] [Medline: [33583200](#)]
3. Greenhouse JB, Stangl D, Bromberg J. An introduction to survival analysis: statistical methods for analysis of clinical trial data. *J Consult Clin Psychol* 1989;57(4):536-544. [doi: [10.1037//0022-006x.57.4.536](#)] [Medline: [2768615](#)]
4. Prinja S, Gupta N, Verma R. Censoring in clinical trials: review of survival analysis techniques. *Indian J Community Med* 2010;35(2):217-221 [FREE Full text] [doi: [10.4103/0970-0218.66859](#)] [Medline: [20922095](#)]
5. Díaz JSP, García Á. Comparison of machine learning models applied on anonymized data with different techniques. : IEEE; 2023 Presented at: 2023 IEEE International Conference on Cyber Security and Resilience (CSR); 31 July 2023 - 02 August 2023; Venice, Italy p. 618-623 URL: <https://ieeexplore.ieee.org/document/10224917> [doi: [10.1109/csr57506.2023.10224917](#)]
6. Antman E. Data sharing in research: benefits and risks for clinicians. *BMJ* 2014;348:g237. [doi: [10.1136/bmj.g237](#)] [Medline: [24458978](#)]

7. Kairouz P, McMahan HB, Avent B, Bellet A, Bennis M, Nitin BA, et al. Advances and open problems in federated learning. In: Foundations and Trends® in Machine Learning. Boston, Massachusetts: Now Foundations and Trends; 2021:1-210.
8. Bonawitz K, Kairouz P, McMahan B, Ramage D. Federated learning and privacy: building privacy-preserving systems for machine learning and data science on decentralized data. *Queueing* 2021;19(5):87-114 [FREE Full text] [doi: [10.1145/3494834.3500240](https://doi.org/10.1145/3494834.3500240)]
9. McMahan B, Ramage D. Federated learning: collaborative machine learning without centralized training data. Google Research. 2017. URL: <https://blog.research.google/2017/04/federated-learning-collaborative.html> [accessed 2024-02-13]
10. McMahan B, Moore E, Ramage D, Hampson S, Arcas BAY. Communication-efficient learning of deep networks from decentralized data. PMLR 2017;54:1273-1282 Singh A, Zhu J, editors.
11. Torkzadehmahani R, Nasirigerdeh R, Blumenthal DB, Kacprowski T, List M, Matschinske J, et al. Privacy-preserving artificial intelligence techniques in biomedicine. *Methods Inf Med* 2022;61(S 01):e12-e27 [FREE Full text] [doi: [10.1055/s-0041-1740630](https://doi.org/10.1055/s-0041-1740630)] [Medline: [35062032](https://pubmed.ncbi.nlm.nih.gov/35062032/)]
12. Brauneck A, Schmalhorst L, Majdabadi MMK, Bakhtiari M, Völker U, Saak CC, et al. Federated machine learning in data-protection-compliant research. *Nat Mach Intell* 2023;5(1):2-4 Springer Science and Business Media LLC. [doi: [10.1038/s42256-022-00601-5](https://doi.org/10.1038/s42256-022-00601-5)]
13. Yang A, Ma Z, Zhang C, Han Y, Hu Z, Zhang W, et al. Review on application progress of federated learning model and security hazard protection. *Digit Commun Netw* 2023;9(1):146-158 [FREE Full text] [doi: [10.1016/j.dcan.2022.11.006](https://doi.org/10.1016/j.dcan.2022.11.006)]
14. Asad M, Moustafa A, Ito T. Federated learning versus classical machine learning: a convergence comparison. ArXiv Preprint posted online on 22 Jul 2021 [FREE Full text] [doi: [10.22541/au.162074596.66890690/v1](https://doi.org/10.22541/au.162074596.66890690/v1)]
15. Sheller MJ, Edwards B, Reina GA, Martin J, Pati S, Kotrotsou A, et al. Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data. *Sci Rep* 2020;10(1):12598 [FREE Full text] [doi: [10.1038/s41598-020-69250-1](https://doi.org/10.1038/s41598-020-69250-1)] [Medline: [32724046](https://pubmed.ncbi.nlm.nih.gov/32724046/)]
16. Zolotareva O, Nasirigerdeh R, Matschinske J, Torkzadehmahani R, Bakhtiari M, Frisch T, et al. Flimma: a federated and privacy-aware tool for differential gene expression analysis. *Genome Biol* 2021;22(1):338 [FREE Full text] [doi: [10.1186/s13059-021-02553-2](https://doi.org/10.1186/s13059-021-02553-2)] [Medline: [34906207](https://pubmed.ncbi.nlm.nih.gov/34906207/)]
17. Nasirigerdeh R, Torkzadehmahani R, Matschinske J, Frisch T, List M, Späth J, et al. sPLINK: a hybrid federated tool as a robust alternative to meta-analysis in genome-wide association studies. *Genome Biol* 2022;23(1):32 [FREE Full text] [doi: [10.1186/s13059-021-02562-1](https://doi.org/10.1186/s13059-021-02562-1)] [Medline: [35073941](https://pubmed.ncbi.nlm.nih.gov/35073941/)]
18. Lu CL, Wang S, Ji Z, Wu Y, Xiong L, Jiang X, et al. WebDISCO: a web service for distributed cox model learning without patient-level data sharing. *J Am Med Inform Assoc* 2015;22(6):1212-1219 [FREE Full text] [doi: [10.1093/jamia/ocv083](https://doi.org/10.1093/jamia/ocv083)] [Medline: [26159465](https://pubmed.ncbi.nlm.nih.gov/26159465/)]
19. Andreux M, Manoel A, Menuet R, Saillard C, Simpson C. Federated survival analysis with discrete-time cox models. ArXiv Preprint posted online on 16 Jun 2020 [FREE Full text]
20. Froelicher D, Troncoso-Pastoriza JR, Raisaro JL, Cuendet MA, Sousa JS, Cho H, et al. Truly privacy-preserving federated analytics for precision medicine with multiparty homomorphic encryption. *Nat Commun* 2021;12(1):5910 [FREE Full text] [doi: [10.1038/s41467-021-25972-y](https://doi.org/10.1038/s41467-021-25972-y)] [Medline: [34635645](https://pubmed.ncbi.nlm.nih.gov/34635645/)]
21. Späth J, Matschinske J, Kamanu FK, Murphy SA, Zolotareva O, Bakhtiari M, et al. Privacy-aware multi-institutional time-to-event studies. *PLOS Digit Health* 2022;1(9):e0000101 [FREE Full text] [doi: [10.1371/journal.pdig.0000101](https://doi.org/10.1371/journal.pdig.0000101)] [Medline: [36812603](https://pubmed.ncbi.nlm.nih.gov/36812603/)]
22. Pölsterl S. scikit-survival: a library for time-to-event analysis built on top of scikit-learn. *J Mach Learn Res* 2020;21(1):8747-8752 [FREE Full text]
23. Matschinske J, Späth J, Bakhtiari M, Probul N, Majdabadi MMK, Nasirigerdeh R, et al. The FeatureCloud platform for federated learning in biomedicine: unified approach. *J Med Internet Res* 2023;25:e42621 [FREE Full text] [doi: [10.2196/42621](https://doi.org/10.2196/42621)] [Medline: [37436815](https://pubmed.ncbi.nlm.nih.gov/37436815/)]
24. Pölsterl S, Navab N, Katouzian A. Fast training of support vector machines for survival analysis. In: Machine Learning and Knowledge Discovery in Databases. Cham, Switzerland: Springer International Publishing; 2015:243-259.
25. FeatureCloud AI Developer API (1.1.0). FeatureCloud. URL: <https://featurecloud.ai/assets/api/redoc-static.html> [accessed 2024-01-13]
26. Cramer R, Damgard IB, Nielsen JB. Secure Multiparty Computation and Secret Sharing. Cambridge, England: Cambridge University Press; 2015.
27. Desmedt C, Piette F, Loi S, Wang Y, Lallemand F, Haibe-Kains B, et al. Strong time dependence of the 76-gene prognostic signature for node-negative breast cancer patients in the TRANSBIG multicenter independent validation series. *Clin Cancer Res* 2007;13(11):3207-3214 [FREE Full text] [doi: [10.1158/1078-0432.CCR-06-2765](https://doi.org/10.1158/1078-0432.CCR-06-2765)] [Medline: [17545524](https://pubmed.ncbi.nlm.nih.gov/17545524/)]
28. Schumacher M, Bastert G, Bojar H, Hübner K, Olschewski M, Sauerbrei W, et al. Randomized 2 x 2 trial evaluating hormonal treatment and the duration of chemotherapy in node-positive breast cancer patients. German Breast Cancer Study Group. *J Clin Oncol* 1994;12(10):2086-2093. [doi: [10.1200/JCO.1994.12.10.2086](https://doi.org/10.1200/JCO.1994.12.10.2086)] [Medline: [7931478](https://pubmed.ncbi.nlm.nih.gov/7931478/)]
29. Hosmer DW, Lemeshow S, May S. Applied Survival Analysis: Regression Modeling of Time to Event Data: Second Edition. New York, NY: John Wiley and Sons Inc; 2008.

30. Solé C, Guilly S, Da Silva K, Llopis M, Le-Chatelier E, Huelin P, et al. Alterations in gut microbiome in cirrhosis as assessed by quantitative metagenomics: relationship with acute-on-chronic liver failure and prognosis. *Gastroenterology* 2021;160(1):206.e13-218.e13 [FREE Full text] [doi: [10.1053/j.gastro.2020.08.054](https://doi.org/10.1053/j.gastro.2020.08.054)] [Medline: [32941879](https://pubmed.ncbi.nlm.nih.gov/32941879/)]
31. Wen C, Zheng Z, Shao T, Liu L, Xie Z, Le Chatelier E, et al. Quantitative metagenomics reveals unique gut microbiome biomarkers in ankylosing spondylitis. *Genome Biol* 2017;18(1):142 [FREE Full text] [doi: [10.1186/s13059-017-1271-6](https://doi.org/10.1186/s13059-017-1271-6)] [Medline: [28750650](https://pubmed.ncbi.nlm.nih.gov/28750650/)]
32. Oñate FP, Le Chatelier E, Almeida M, Cervino ACL, Gauthier F, Magoulès F, et al. MSPminer: abundance-based reconstitution of microbial pan-genomes from shotgun metagenomic data. *Bioinformatics* 2019;35(9):1544-1552 [FREE Full text] [doi: [10.1093/bioinformatics/bty830](https://doi.org/10.1093/bioinformatics/bty830)] [Medline: [30252023](https://pubmed.ncbi.nlm.nih.gov/30252023/)]
33. Harrell FE, Califf RM, Pryor DB, Lee KL, Rosati RA. Evaluating the yield of medical tests. *JAMA* 1982;247(18):2543-2546. [Medline: [7069920](https://pubmed.ncbi.nlm.nih.gov/7069920/)]
34. Dwork C. Differential privacy. In: *Automata, Languages and Programming*. Berlin, Heidelberg: Springer; 2006:1-12.
35. Yeom S, Giacomelli I, Fredrikson M, Jha S. Privacy risk in machine learning: analyzing the connection to overfitting. : *IEEE*; 2018 Presented at: 2018 IEEE 31st Computer Security Foundations Symposium (CSF); July 09-12, 2018; Oxford, UK p. 268-282 URL: <https://ieeexplore.ieee.org/abstract/document/8429311/> [doi: [10.1109/csf.2018.00027](https://doi.org/10.1109/csf.2018.00027)]
36. Hsu J, Gaboardi M, Haebleren A, Khanna S, Narayan A, Pierce BC, et al. Differential privacy: an economic method for choosing epsilon. : *IEEE*; 2014 Presented at: 2014 IEEE 27th Computer Security Foundations Symposium; July 19-22, 2014; Vienna, Austria p. 398-410 URL: <https://ieeexplore.ieee.org/abstract/document/6957125/> [doi: [10.1109/csf.2014.35](https://doi.org/10.1109/csf.2014.35)]
37. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. USA: Curran Associates Inc; 2017 Presented at: Proceedings of the 31st International Conference on Neural Information Processing Systems Red Hook; 2017; NY, USA p. 4768-4777 URL: https://proceedings.neurips.cc/paper_files/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf
38. Smith A, Lambert PC, Rutherford MJ. Generating high-fidelity synthetic time-to-event datasets to improve data transparency and accessibility. *BMC Med Res Methodol* 2022;22(1):176 [FREE Full text] [doi: [10.1186/s12874-022-01654-1](https://doi.org/10.1186/s12874-022-01654-1)] [Medline: [35739465](https://pubmed.ncbi.nlm.nih.gov/35739465/)]
39. Merkel D. Docker: Lightweight Linux Containers for Consistent Development and Deployment. Linux J Houston, TX: Belltown Media; 2014. URL: <https://www.seltzer.com/margo/teaching/CS508.19/papers/merkel14.pdf> [accessed 2024-03-06]
40. Kaplan EL, Meier P. Nonparametric estimation from incomplete observations. *J Am Stat Assoc* 1958;53(282):457-481. [doi: [10.1080/01621459.1958.10501452](https://doi.org/10.1080/01621459.1958.10501452)]
41. Aalen O. Nonparametric inference for a family of counting processes. *Ann Statist* 1978;6(4):701-726. [doi: [10.1214/aos/1176344247](https://doi.org/10.1214/aos/1176344247)]
42. Cox D. Regression models and life-tables. *J R Stat Soc* 1972;34(2):187-202 [FREE Full text] [doi: [10.1007/978-1-4612-4380-9_37](https://doi.org/10.1007/978-1-4612-4380-9_37)]
43. Ishwaran H, Kogalur UB, Blackstone EH, Lauer MS. Random survival forests. *Ann Appl Stat* 2008;2(3):841-860. [doi: [10.1214/08-aos169](https://doi.org/10.1214/08-aos169)]
44. Hauschild AC, Lemanczyk M, Matschinske J, Frisch T, Zolotareva O, Holzinger A, et al. Federated random forests can improve local performance of predictive models for various healthcare applications. *Bioinformatics* 2022;38(8):2278-2286 [FREE Full text] [doi: [10.1093/bioinformatics/btac065](https://doi.org/10.1093/bioinformatics/btac065)] [Medline: [35139148](https://pubmed.ncbi.nlm.nih.gov/35139148/)]
45. Shamir A. How to share a secret. *Commun ACM* 1979;22(11):612-613 [FREE Full text] [doi: [10.1145/359168.359176](https://doi.org/10.1145/359168.359176)]
46. Kairouz P, Brendan MH, Avent B, Bellet A, Bennis M, Bhagoji AN, et al. Advances and open problems in federated learning. *ArXiv Preprint posted online on 9 Mar 2021* [FREE Full text] [doi: [10.1561/9781680837896](https://doi.org/10.1561/9781680837896)]
47. Liu P, Xu X, Wang W. Threats, attacks and defenses to federated learning: issues, taxonomy and perspectives. *Cybersecurity* 2022;5(1):1-19 [FREE Full text] [doi: [10.1186/s42400-021-00105-6](https://doi.org/10.1186/s42400-021-00105-6)]
48. Normalization app. FeatureCloud. 2022. URL: <https://featurecloud.ai/app/normalization> [accessed 2024-01-13]
49. Cross validation app. FeatureCloud. 2022. URL: <https://featurecloud.ai/app/cross-validation> [accessed 2024-01-13]
50. Späth J. julianspaeth / federated-survival-svm. GitHub. URL: <https://github.com/julianspaeth/federated-survival-svm> [accessed 2024-03-21]
51. Späth J. FeatureCloud / fc-survival-svm. GitHub. URL: <https://github.com/FeatureCloud/fc-survival-svm> [accessed 2024-03-21]

Abbreviations

- BRCA:** breast cancer data set
- CV:** cross-validation
- DP:** differential privacy
- FL:** federated learning
- GBSG2:** German Breast Cancer Study Group 2 data set
- GDPR:** General Data Protection Regulation
- HE:** homomorphic encryption

ML: machine learning
MSP: Metagenomic Species Pangenome
PET: privacy-enhancing technique
RSA: Rivest–Shamir–Adleman
SHAP: Shapley additive explanations
SMPC: secure multiparty computation
SVM: support vector machine
WHAS500: Worcester Heart Attack Study data set

Edited by K El Emam, B Malin; submitted 30.03.23; peer-reviewed by N Mungoli, S Nagavally, R Gorantla, D Gopukumar, X Jiang, Y Huang; comments to author 02.07.23; revised version received 06.08.23; accepted 10.02.24; published 29.03.24.

Please cite as:

*Späth J, Sewald Z, Probul N, Berland M, Almeida M, Pons N, Le Chatelier E, Ginès P, Solé C, Juanola A, Pauling J, Baumbach J
Privacy-Preserving Federated Survival Support Vector Machines for Cross-Institutional Time-To-Event Analysis: Algorithm Development and Validation*

JMIR AI 2024;3:e47652

URL: <https://ai.jmir.org/2024/1/e47652>

doi: [10.2196/47652](https://doi.org/10.2196/47652)

PMID: [38875678](https://pubmed.ncbi.nlm.nih.gov/38875678/)

©Julian Späth, Zeno Sewald, Niklas Probul, Magali Berland, Mathieu Almeida, Nicolas Pons, Emmanuelle Le Chatelier, Pere Ginès, Cristina Solé, Adrià Juanola, Josch Pauling, Jan Baumbach. Originally published in JMIR AI (<https://ai.jmir.org>), 29.03.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR AI, is properly cited. The complete bibliographic information, a link to the original publication on <https://www.ai.jmir.org/>, as well as this copyright and license information must be included.

Original Paper

Machine Learning–Based Prediction for High Health Care Utilizers by Using a Multi-Institutional Diabetes Registry: Model Training and Evaluation

Joshua Kuan Tan¹, MBBS, MScPH; Le Quan², MSc; Nur Nasyitah Mohamed Salim¹, BSc; Jen Hong Tan², PhD; Su-Yen Goh³, MBBS; Julian Thumboo¹, MMED; Yong Mong Bee³, MBBS

¹Health Services Research Unit, Singapore General Hospital, Singapore, Singapore

²Data Science and Artificial Intelligence Laboratory, Singapore General Hospital, Singapore, Singapore

³Department of Endocrinology, Singapore General Hospital, Singapore, Singapore

Corresponding Author:

Joshua Kuan Tan, MBBS, MScPH

Health Services Research Unit

Singapore General Hospital

10 Hospital Blvd

Singapore, 168582

Singapore

Phone: 65 6222 3322

Email: joshua.tank@mohh.com.sg

Abstract

Background: The cost of health care in many countries is increasing rapidly. There is a growing interest in using machine learning for predicting high health care utilizers for population health initiatives. Previous studies have focused on individuals who contribute to the highest financial burden. However, this group is small and represents a limited opportunity for long-term cost reduction.

Objective: We developed a collection of models that predict future health care utilization at various thresholds.

Methods: We utilized data from a multi-institutional diabetes database from the year 2019 to develop binary classification models. These models predict health care utilization in the subsequent year across 6 different outcomes: patients having a length of stay of ≥ 7 , ≥ 14 , and ≥ 30 days and emergency department attendance of ≥ 3 , ≥ 5 , and ≥ 10 visits. To address class imbalance, random and synthetic minority oversampling techniques were employed. The models were then applied to unseen data from 2020 and 2021 to predict health care utilization in the following year. A portfolio of performance metrics, with priority on area under the receiver operating characteristic curve, sensitivity, and positive predictive value, was used for comparison. Explainability analyses were conducted on the best performing models.

Results: When trained with random oversampling, 4 models, that is, logistic regression, multivariate adaptive regression splines, boosted trees, and multilayer perceptron consistently achieved high area under the receiver operating characteristic curve (>0.80) and sensitivity (>0.60) across training-validation and test data sets. Correcting for class imbalance proved critical for model performance. Important predictors for all outcomes included age, number of emergency department visits in the present year, chronic kidney disease stage, inpatient bed days in the present year, and mean hemoglobin A_{1c} levels. Explainability analyses using partial dependence plots demonstrated that for the best performing models, the learned patterns were consistent with real-world knowledge, thereby supporting the validity of the models.

Conclusions: We successfully developed machine learning models capable of predicting high service level utilization with strong performance and valid explainability. These models can be integrated into wider diabetes-related population health initiatives.

(JMIR AI 2024;3:e58463) doi:[10.2196/58463](https://doi.org/10.2196/58463)

KEYWORDS

diabetes mellitus; type 2 diabetes; health care utilization; population health management; population health; machine learning; artificial intelligence; predictive model; predictive system; practical model

Introduction

In recent years, high-income countries worldwide have seen a consistent rise in health care expenditure. Singapore, mirroring this trend, has experienced a steady increase in health care spending relative to its gross domestic product [1]. To address this, Singapore is undergoing a transformative health system initiative known as Healthier SG [2], which is an initiative to pivot the health system toward preventive care and population health management.

Parallel to these efforts, there is a burgeoning interest in leveraging machine learning for individual-level health utilization predictions. Identifying prospective high utilizers of health care services could unlock opportunities for targeted interventions. These interventions are poised not only to enhance individual health outcomes but also to reduce long-term health care utilization and system costs. Existing research suggests that a disproportionate amount of health care spending is concentrated among a small group of costly patients known as the high-need, high-cost (HNHC) patients—often defined as those who account for the top 5% of the annual health care costs [3,4]. These patients were believed to present an opportunity for cost reduction [5].

However, the potential for cost savings in caring for HNHC patients is often less than anticipated [6]. This is due to the diverse nature of these patients who can be subdivided into 3 categories: persistent and refractory HNHC patients, individuals who experience a 1-time catastrophic health event, and patients with multiple chronic conditions but amenable to disease management programs [6,7]. Notably, the latter group presents the most viable opportunity for impactful intervention. Persistent and refractory HNHC patients are those with severe and chronic diseases who require ongoing and expensive care. For these patients, disease management programs often do not result in significant reduction in health utilization and financial savings. For patients with 1-time catastrophic health events such as accidents, these events are difficult to predict and therefore not amenable to any intervention [6,7]. Therefore, targeting the small cohort with multiple chronic conditions but amenable to disease management programs represents a limited opportunity to reduce health care costs [6].

Given these complexities, there is a need to refine the approach to predicting and managing high health care utilization. One strategy could be to expand the predictive scope beyond HNHC patients or explore other indicators. Relatedly, the total length of stay (LOS) and frequency of emergency department (ED) visits per calendar year may provide a better indication of service-related health care utilization and the intensity of inpatient resource use [8].

This study aims to develop prediction models to forecast annual inpatient bed days and ED utilization across varying thresholds; presently, such models are not available in our hospital system. We utilized the Singapore Health Services (SingHealth) Diabetes Registry (SDR), a comprehensive clinical database of patients with diabetes within our hospital system to develop predictive models. Our objective is to create clinically relevant and

actionable models that can be integrated into wider diabetes-related population health initiatives [9].

Methods

Study Setting

We used data from the multi-institutional SDR, previously described in detail [10]. SingHealth is the largest of the 3 public health care clusters in Singapore and manages 4 acute hospitals, 5 national specialty centers, 3 community hospitals, and a network of 10 primary care polyclinics. SDR was initiated in 2015 and populated retrospectively and prospectively from across SingHealth's electronic medical records and clinical databases to cover the period of 2013 to 2022.

Outcome Variables

As SDR primarily consists of clinical data from electronic medical records and lacks financial information, we focused on service-related health care utilization metrics. To this end, we developed models to predict utilization across 6 different thresholds (per calendar year), specifically for total LOS at ≥ 7 , ≥ 14 , and ≥ 30 days and for ED attendance ≥ 3 , ≥ 5 , and ≥ 10 visits; thus, 6 sets of (binary classification) models were constructed. Currently, there are no standard definitions for long inpatient LOS or high ED attendance.

For total LOS, we set arbitrary thresholds corresponding to 1 week, 2 weeks, and 1 month. These thresholds were chosen to reflect varying degrees of health care utilization in ours and possibly other health care systems, corresponding to different levels of patient care needs and resource allocation. Inpatient stays between 1 and 2 weeks represent short-term stays, potentially indicative of acute or less severe conditions. In contrast, stays longer than 2 weeks and those extending beyond 1 month represent increasingly prolonged stays, often associated with more severe or complex health issues, especially in the latter. These distinctions are critical for understanding and managing different patient care strategies. They also represent varying levels of health care management and resource planning, as we intend to develop disease management programs around these thresholds in the future. Regarding ED attendance, a recent systematic review indicated that ≥ 3 was the most common definition for high ED attendance but noted that definitions could extend to 30 or more visits [11]. Accordingly, we defined high ED attendance by using the 3 aforementioned thresholds, with ≥ 3 visits as the minimum criterion. This approach may aid in planning interventions to prevent escalation to higher levels of utilization.

Explanatory Variables

The SDR data set facilitated an examination of the effects of sociodemographic indicators, health indicators, and diabetes-related complications. Our methodology for ascertaining diabetes-related complications has been published previously [12] and detailed in Table S1 of [Multimedia Appendix 1](#). The models incorporated 24 variables detailed in Table S2 of [Multimedia Appendix 1](#). These variables are readily derived from electronic medical records during admissions, ED visits, inpatient and outpatient clinical consultations, and are based on local clinical guidelines [13]. These variables offer a

comprehensive view of the patients from demographic, social, clinical, and utilization perspectives.

Inclusion and Exclusion Criteria

This study utilizes data from SDR spanning 2019 to 2022, as this was the period when comprehensive health care utilization data were available. We included patients aged 18 years and older diagnosed with type 2 diabetes mellitus. Patients with missing variables were excluded from this study, as we did not perform data imputation, and most machine learning algorithms do not support missing values.

Handling Unbalanced Data

Our data set demonstrated significant class imbalance in inpatient and ED utilization, which can bias models toward the majority class, hinder the identification of the high utilizers (the minority class) [14], and result in subpar model performance. In this study, we utilized oversampling, a data-level method to address the class imbalance. Specifically, we used the synthetic minority oversampling technique-nominal continuous (SMOTE-NC) [15] from the *themis* package [16]. SMOTE-NC, a variant of the SMOTE family of algorithms, generates new examples of the minority class by interpolating between several minority class instances that lie relatively close to each other [17]. SMOTE-NC is effective with mixed numerical and categorical data. We applied SMOTE-NC with $k=5$ and $k=3$ settings, where k denotes the number of nearest neighbors used to generate new examples of the minority class. Additionally, we used the *upSample* algorithm from the *caret* package [18] for random oversampling and compared it with no oversampling. All oversampling techniques achieved equal representations of both classes in our training data set (ie, equal number of patients with and without the outcome in the training data set).

Performance Indicators

We assessed model performance by using area under the receiver operating characteristic curve (AUC), sensitivity (recall), and positive predictive value (PPV). Sensitivity (recall) allowed us to identify whether the models were able to correctly identify patients with the outcomes of interest. PPV provided us with an understanding of the quality of the positive prediction made by the model. Additionally, we have reported the area under the precision-recall curve, sensitivity, specificity, and F_1 -score in [Multimedia Appendix 1](#). The area under the precision-recall curve is preferred over AUC for rare outcomes, as it more accurately reflects model performance [19]. We also evaluated the confusion matrix during model development.

Machine Learning Models

We built 7 predictive models using R software (version 4.3.1; R Foundation for Statistical Computing) and the *tidymodels*

package [20]: logistic regression, random forest, boosted trees, multilayer perceptron (MLP), k-nearest neighbor, multivariate adaptive regression splines (MARS), and Bayesian additive regression trees. SDR data from 2019 were randomly split into training (75%) and validation (25%) data sets, with no overlap between the data sets. Since the training data set was large ($n=75,375$), we did not perform cross-validation during model training. No hyperparameter tuning was performed, as the intent of the study was to build baseline models to understand the problem and data set while prioritizing model simplicity and interpretability. The trained models were then tested on unseen data from 2020 and 2021 (ie, the model utilized 2020 data to predict 2021 outcomes and 2021 data to predict 2022 outcomes). Although the data sets originate from the same registry, they reflect distinct utilization patterns across different years, ensuring temporal independence between them.

Explainability

For top-performing models, model interpretation was determined using model-specific variable importance scores with the *vip* package [21] and permutation feature importance plots using the *DALEX* package [22,23]. Additionally, for the top variables identified through these methods, partial dependence plots (PDPs) were generated using the *DALEX* package and the unseen validation data set to visualize the relationship between key predictor variables and the probability of the outcome occurring.

Ethics Approval

Ethics approval was obtained from the SingHealth Centralized Institutional Review Board prior to initiating this study (reference: 2022/2133). As all participant data were deidentified, a waiver for participant consent was also obtained.

Reporting Checklist

We followed the consolidated reporting guidelines for prognostic and diagnostic machine learning modeling studies [24] (Table S3 in [Multimedia Appendix 1](#)).

Results

Characteristics of the Data Sets

After removing patients with missing data from the registry in 2019, the training data set contained 100,500 (74.6%) individuals of the 134,670 patients in SDR in 2019. The test sets in 2020 and 2021 comprised 77.3% (108,886/140,859) and 80.7% (111,004/137,584) of the total SDR cohorts for the respective years. The characteristics of the patients included in the training-validation and 2 test data sets are described in detail in [Table 1](#).

Table 1. Demographics, comorbidities, and utilization characteristics of the training and test data sets.

Data set description	Training and validation ^a 2019-2020 (n=134,670)	Test 2020-2021 (n=140,859)	Test 2021-2022 (n=137,584)
Data set size, n (% of total registry)	100,500 (74.6)	108,886 (77.3)	111,004 (80.7)
Female gender, n (%)	48,887 (48.6)	52,210 (48)	53,148 (47.9)
Age on January 1 at the start of the year (years)			
Mean (SD)	66.4 (11.8)	66.7 (11.9)	66.5 (12.2)
Median	67	67	67
Ethnicity, n (%)			
Chinese	71,132 (70.8)	76,479 (70.2)	76,627 (69)
Malay	14,903 (14.8)	16,277 (15)	17,144 (15.4)
Indian	10,119 (10.1)	11,267 (10.4)	11,788 (10.6)
Other	4346 (4.3)	4863 (4.5)	5445 (4.9)
Housing type, n (%)			
1- and 2-room public housing	7502 (7.5)	8214 (7.5)	10,086 (9.1)
3-room public housing	24,976 (24.9)	26,741 (24.6)	24,779 (22.3)
4-room public housing	32,089 (31.9)	34,933 (32.1)	36,540 (32.9)
5-room public housing and executive flats	25,769 (25.6)	27,942 (25.7)	29,220 (26.3)
Private condominium	6268 (6.2)	6843 (6.3)	6607 (6)
Private landed housing	3896 (3.9)	4213 (3.9)	3772 (3.4)
Lives in a rental block	6641 (6.6)	7290 (6.7)	7294 (6.6)
Comorbidities, n (%)			
Hypertension	87,931 (87.5)	97,149 (89.2)	99,597 (89.7)
Hyperlipidemia	95,679 (95.2)	105,108 (96.5)	107,638 (97)
Diabetes mellitus medications, n (%)			
None	18,125 (18)	20,712 (19)	18,426 (16.6)
Oral medications only	57,413 (57.1)	64,571 (59.3)	61,516 (55.4)
Insulin only	2809 (2.8)	2264 (2.1)	3216 (2.9)
Oral and insulin	22,153 (22)	21,339 (19.6)	27,846 (25.1)
Diabetes-related complications, n (%)			
Ischemic heart disease	25,097 (25)	27,663 (25.4)	30,656 (27.6)
Ischemic stroke	9401 (9.4)	10,563 (9.7)	11,305 (10.2)
Hemorrhagic stroke	1449 (1.4)	1801 (1.7)	1998 (1.8)
Peripheral arterial disease	3910 (3.9)	4577 (4.2)	5198 (4.7)
Major lower-extremity amputation	138 (0.1)	173 (0.2)	182 (0.2)
Minor lower-extremity amputation	339 (0.3)	340 (0.3)	426 (0.4)
Diabetic foot and peripheral angiopathy	2718 (2.7)	3180 (2.9)	3524 (3.2)
Diabetic eye complications	13,067 (13)	13,116 (12.1)	14,479 (13)
Nephropathy	49,139 (48.9)	53,737 (49.4)	54,359 (49)
Chronic kidney disease stage, n (%)			
1 (eGFR ^b ≥90)	35,176 (35)	36,603 (33.6)	37,188 (33.5)
2 (eGFR 60-89)	41,705 (41.5)	45,216 (41.5)	45,755 (41.2)

Data set description	Training and validation ^a 2019-2020 (n=134,670)	Test 2020-2021 (n=140,859)	Test 2021-2022 (n=137,584)
3A (eGFR 45-59)	11,563 (11.5)	12,667 (11.6)	12,802 (11.5)
3B (eGFR 30-44)	6760 (6.7)	7696 (7.1)	7835 (7.1)
4 (eGFR 15-29)	3215 (3.2)	3805 (3.5)	4016 (3.6)
5 (eGFR<15)	2081 (2.1)	2899 (2.7)	3408 (3.1)
Dialysis	1400 (1.4)	1903 (1.8)	2269 (2)
Utilization characteristics			
Inpatient utilization (present year)			
Mean (SD)	3.09 (11.3)	3.41 (11.7)	3.96 (13.6)
Median	0	0	0
Inpatient bed days (present year), n (%)			
0	77,170 (76.8)	81,559 (74.9)	80,770 (72.8)
1-2	6034 (6)	6752 (6.2)	7168 (6.5)
3-6	6693 (6.7)	7701 (7.1)	8500 (7.7)
7-13	4464 (4.4)	5432 (5)	5982 (5.4)
14-29	3592 (3.6)	4315 (4)	4855 (4.4)
≥30	2547 (2.5)	3127 (2.9)	3729 (3.4)
Inpatient bed days (subsequent year)			
Mean (SD)	2.39 (10.3)	2.79 (12.2)	3.22 (14)
Median	0	0	0
Inpatient bed days category (subsequent year), n (%)			
0	83,759 (83.3)	90,022 (82.7)	89,577 (80.7)
1-2	4078 (4.1)	4214 (3.9)	4561 (4.1)
3-6	4477 (4.5)	5015 (4.6)	5619 (5.1)
7-13	3353 (3.3)	3729 (3.4)	4292 (3.9)
14-29	2740 (2.7)	3222 (3)	3722 (3.4)
≥30	2093 (2.1)	2684 (2.5)	3233 (2.9)
Emergency department utilization (present year)			
Mean (SD)	0.53 (1.4)	0.54 (1.4)	0.57 (1.6)
Median	0	0	0
Emergency department visit category (present year), n (%)			
0 visits	71,584 (71.2)	76,261 (70)	75,376 (67.9)
1-2 visits	23,487 (23.4)	27,143 (24.9)	29,671 (26.7)
3-4 visits	3883 (3.9)	3938 (3.6)	4343 (3.9)
5-9 visits	1348 (1.3)	1358 (1.3)	1403 (1.3)
≥10 visits	198 (0.2)	186 (0.2)	211 (0.2)
Emergency department utilization (subsequent year)			
Mean (SD)	0.40 (1.3)	0.40 (1.4)	0.48 (1.4)
Median	0	0	0
Emergency department visit category (subsequent year), n (%)			
0 visits	78,849 (78.5)	85,162 (78.2)	82,269 (74.1)
1-2 visits	17,794 (17.7)	19,434 (17.9)	23,273 (21)
3-4 visits	2716 (2.7)	3060 (2.8)	3817 (3.4)

Data set description	Training and validation ^a 2019-2020 (n=134,670)	Test 2020-2021 (n=140,859)	Test 2021-2022 (n=137,584)
5-9 visits	996 (1)	1064 (1)	1455 (1.3)
≥10 visits	145 (0.1)	166 (0.2)	190 (0.2)

^aThe data set was randomly partitioned into training and validation data sets in a 75% to 25% ratio (respectively), with no overlap between the 2 data sets. n=total registry size.

^beGFR: estimated glomerular filtration rate in mL/min/1.73 m².

Across the data sets, 47.9%-48.6% of the patients were females. The mean age was between 66.4 and 66.7 years, and the median was consistently 67 years. The proportions by ethnicities were consistent across the 3 data sets with approximately 70% Chinese, 14% Malay, 10% Indian, and 4% other races. The ethnic distributions observed closely resembled the Singaporean population [25]. Across the data sets, most individuals lived in public housing, with the largest proportion being 4-room public housing (approximately 32%). Owing to the public housing infrastructure in Singapore, approximately 6.6% of the patients live in an apartment block with rental housing. Across the data sets, the proportion of patients with hypertension was 87.5%-89.7%, whereas the proportion of patients with hyperlipidemia was 95.2%-97%. The most common diabetes-related complication was nephropathy (prevalence of 48.9%-49.4% across the data sets) followed by ischemic heart disease (prevalence of 25%-27.6%) and then diabetic eye complications (prevalence of 12.1%-13%). Relatedly, 65%-66.5% of the patients in the data sets had stage 2 chronic kidney disease (CKD) and above. When contrasted with the prevalence of nephropathy (our definition of nephropathy was estimated glomerular filtration rate <60 mL/min/1.73 m² or urine albumin creatinine ratio ≥30 mg/g or urine protein/creatinine ratio ≥0.20 g/g), it suggests that a significant proportion of patients had stage 1 CKD and proteinuria.

The mean present year inpatient utilization across the data sets was 3.08%-3.96%. Compared to the present year, the subsequent

year's inpatient utilization was less. The mean present year ED utilization was 0.53-0.57 visits per patient. Compared to the present year, the subsequent year's ED utilization was less. The median utilization for present and next year's inpatient and ED utilization was zero across all data sets, indicating that the utilization characteristics were extremely skewed.

Effects of Sampling Technique on Model Performance

The key model performance indices for the models using different oversampling techniques and no oversampling are presented in Figures 1-2 (Figures 1-2 in Multimedia Appendix 2) and Table S4 in Multimedia Appendix 1. For all the outcomes studied, models trained with random oversampling had similar AUC values to models trained with no oversampling, models trained with SMOTE-NC (k=3) had lower AUC values, and models trained with SMOTE-NC (k=5) had the lowest AUC. With regard to sensitivity, models trained with no oversampling had markedly lower sensitivity but higher PPVs. This indicates that models trained with no oversampling could not correctly identify patients with the outcomes of interest. This is further confirmed in our analysis of the confusion matrixes of these models trained. We observed that these models assigned almost all the patients as not cases (ie, did not have the outcomes the next year) and therefore were not useful. Models trained with no oversampling and SMOTE-NC (k=5) were not included in further analyses.

Figure 1. Comparing between different oversampling techniques to predict inpatient bed days. A. Predicting ≥ 7 inpatient bed days in subsequent year. B. Predicting ≥ 14 inpatient bed days in subsequent year. C. Predicting ≥ 30 inpatient bed days in subsequent year. AUC: area under the receiver operating characteristic curve; BART: Bayesian additive regression trees; KNN: k-nearest neighbor; MARS: multivariate adaptive regression splines; MLP: multilayer perceptron; PPV: positive predictive value; SMOTE-NC: synthetic minority oversampling technique-nominal continuous. A higher-resolution image of this figure is available in [Multimedia Appendix 2](#).

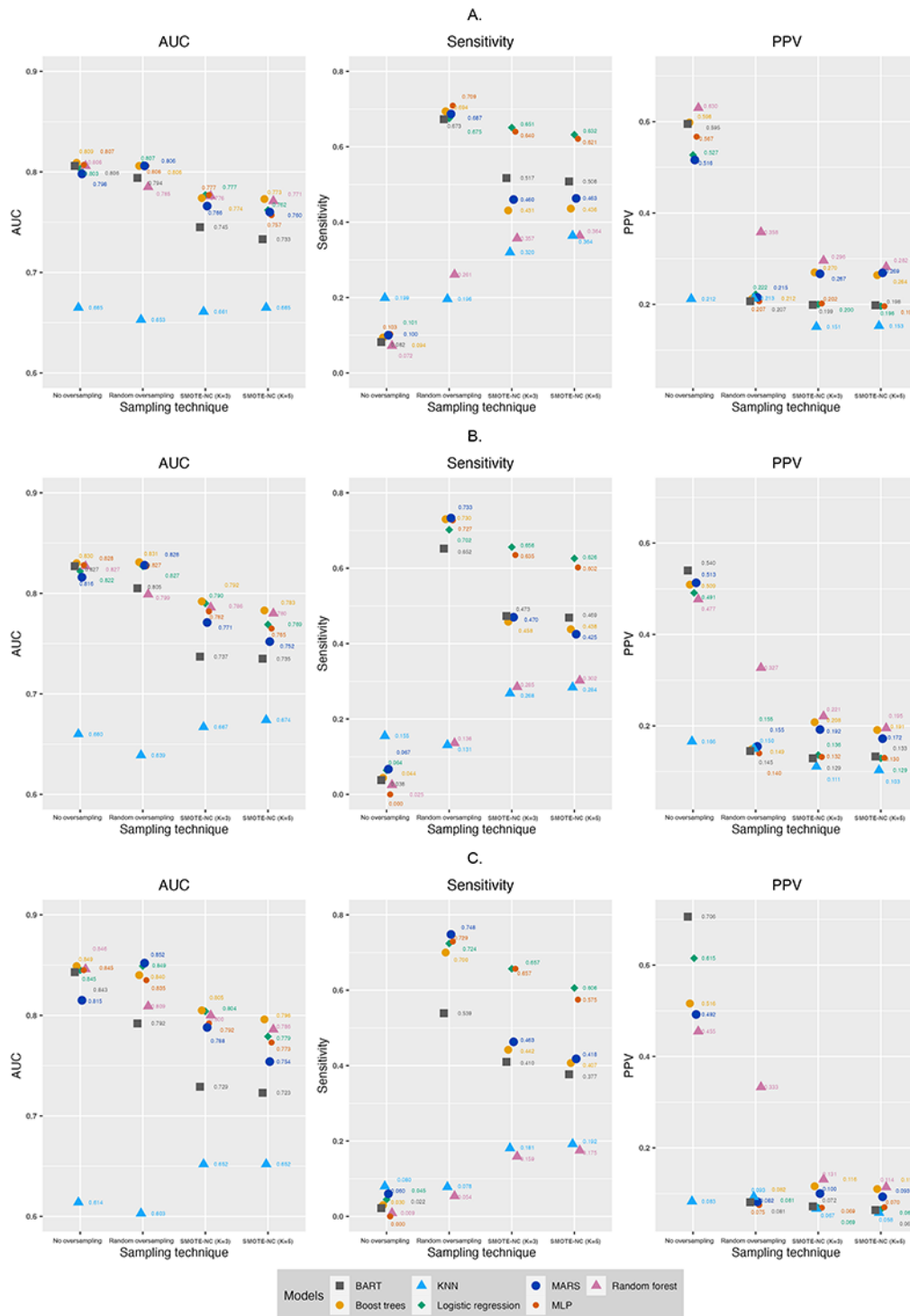


Figure 2. Comparing between different oversampling techniques to predict emergency department visits. A. Predicting ≥ 3 emergency department visits in subsequent year. B. Predicting ≥ 5 emergency department visits in subsequent year. C. Predicting ≥ 10 emergency department visits in subsequent year. AUC: area under the receiver operating characteristic curve; BART: Bayesian additive regression trees; KNN: k-nearest neighbor; MARS: multivariate adaptive regression splines; MLP: multilayer perceptron; PPV: positive predictive value; SMOTE-NC: synthetic minority oversampling technique-nominal continuous. A higher-resolution image of this figure is available in [Multimedia Appendix 2](#).



Model Performance on Test Data Sets

As models trained with random oversampling and SMOTE-NC, where $k=3$ had the best AUC and sensitivity, we conducted additional analyses to evaluate their performance by testing them on 2 test data sets of 2020-2021 and 2021-2022 (Figures

3-4, Figures 3-4 in [Multimedia Appendix 2](#), Figures S1-S2 and Tables S5-S6 in [Multimedia Appendix 1](#)). When trained with random oversampling, 4 models, that is, logistic regression, MARS, boosted trees, and MLP had consistently high AUCs across validation and test data sets. The AUC values were higher for outcomes reflecting higher utilization (ie, ≥ 30 inpatient bed

days and ≥ 10 ED visits in subsequent year). These 4 models consistently had the highest sensitivity values, with sensitivity > 0.65 for all outcomes except predicting ≥ 10 ED visits in the subsequent year. This suggests that these 4 models were able to correctly identify at least 65% of the patients with the outcome. All models, except for random forest, had similar but low PPVs across the 2 data sets.

When trained with SMOTE-NC ($k=3$), most models except for k -nearest neighbor and Bayesian additive regression trees models had good AUC (> 0.75) across the 2 test data sets. Models had higher AUC values for outcomes reflecting higher utilization, that is, ≥ 30 inpatient bed days and ≥ 10 ED visits in the subsequent year. Compared to models trained with random oversampling, models trained with SMOTE-NC ($k=3$) had a

wide distribution of sensitivity values, with logistic regression and MLP having similar and consistently high sensitivity values for all outcomes except predicting ≥ 10 ED visits in the subsequent year. Models trained with SMOTE-NC ($k=3$) had a wider distribution of PPV values than models trained with random oversampling.

When comparing the performance of models trained with the 2 oversampling techniques, we observed that random oversampling resulted in marginally higher AUC and sensitivity values (Figures 3-4). The narrow distribution of PPV values in models trained with random oversampling suggests that random oversampling resulted in more consistent quality of positive predictions across the best performing models.

Figure 3. Performance of models trained using random oversampling to predict inpatient bed days. A. Predicting ≥ 7 inpatient bed days in subsequent year. B. Predicting ≥ 14 inpatient bed days in subsequent year. C. Predicting ≥ 30 inpatient bed days in subsequent year. AUC: area under the receiver operating characteristic curve; BART: Bayesian additive regression trees; KNN: k-nearest neighbor; MARS: multivariate adaptive regression splines; MLP: multilayer perceptron; PPV: positive predictive value. A higher-resolution image of this figure is available in [Multimedia Appendix 2](#).

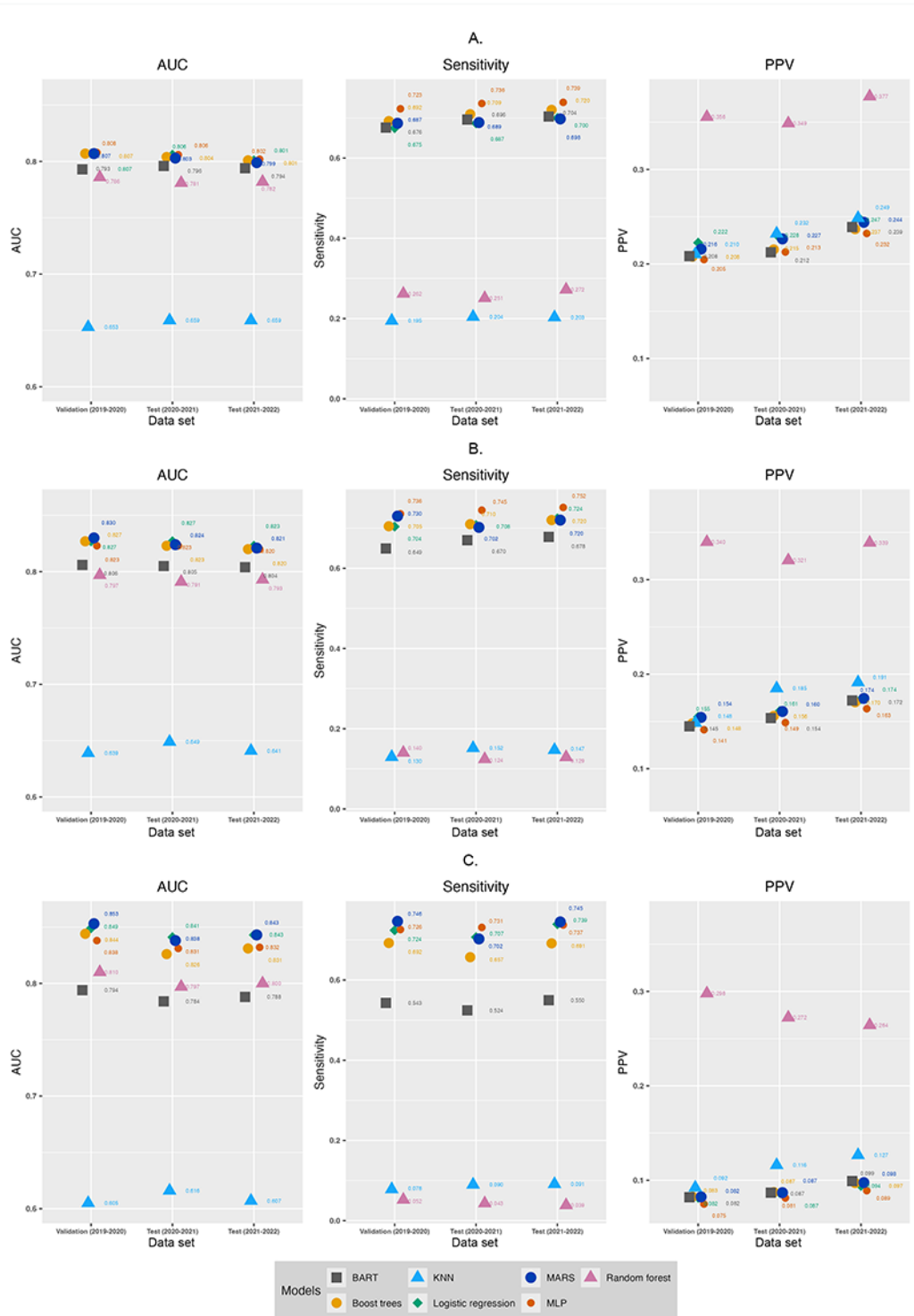
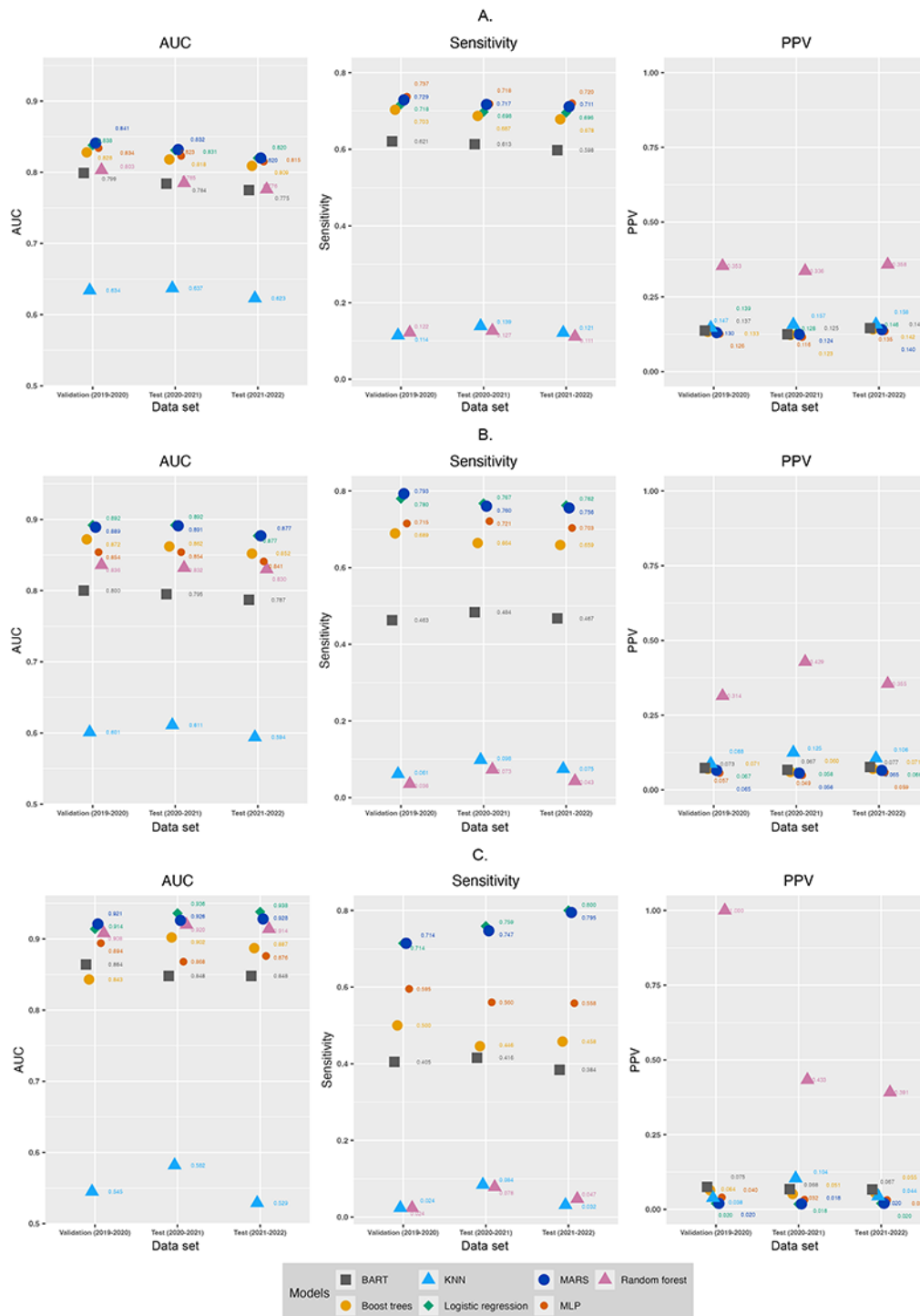


Figure 4. Performance of models trained using random oversampling to predict emergency department visits. A. Predicting ≥ 3 emergency department visits in subsequent year. B. Predicting ≥ 5 emergency department visits in subsequent year. C. Predicting ≥ 10 emergency department visits in subsequent year. A higher resolution version of this figure is available in [Multimedia Appendix 2](#). AUC: area under the receiver operating characteristic curve; BART: Bayesian additive regression trees; KNN: k-nearest neighbor; MARS: multivariate adaptive regression splines; MLP: multilayer perceptron; PPV: positive predictive value.



Explainability Analyses

From our analysis, the best performing models were logistic regression, MARS, boosted trees, and MLP that were trained with random oversampling (herein referred to as selected models). Model-specific variable importance scores for selected

models except MLP were obtained; the top 10 variables are reported in Table S7 in [Multimedia Appendix 1](#). Model-specific variable importance scores for MLP were not available through the *vip* package. Regarding the prediction of subsequent year inpatient bed days (≥ 7 , ≥ 14 , ≥ 30), age, number of ED visits (present year), CKD stages 4 and 5, and present year inpatient

utilization were the most important variables. For boosted tree and MARS, the number of ED visits (present year), CKD stage, and age were the most important variables. Regarding the prediction of subsequent year ED visits, the number of ED visits (present year), CKD stage 4 and 5, mean hemoglobin A_{1c} (HbA_{1c}) values, and age were the most important variables for all models. Interestingly, the number of ED visits (present year) was consistently the most important variable for all the models.

We also obtained permutation feature importance plots for selected models (Figures S3-S4 in [Multimedia Appendix 1](#)). Regarding the prediction of subsequent year inpatient bed days (≥ 7 , ≥ 14 , ≥ 30), the permutation feature importance plots corroborated the model-specific variable importance scores, indicating that age, number of ED visits (present year), CKD stage, and present year inpatient utilization were the most important variables. Interestingly, diabetes mellitus medication category was more important in predicting ≥ 30 inpatient bed days in the subsequent year. Regarding the prediction of subsequent year ED visits, the number of ED visits (present year) was the dominant variable for all models. Other important variables included age, CKD stage, and present year inpatient utilization.

PDPs for the 8 most important variables across selected models are illustrated in [Multimedia Appendix 1](#). Regarding the prediction of inpatient bed days (Figures S5-S7 in [Multimedia Appendix 1](#)), the average prediction of outcomes increased steadily with age for all models. For present-year ED visits, all models demonstrated a sharp increase in average prediction from 0 to 20 visits, with a plateau close to 1.0 (for average prediction) after 20 visits. For present-year inpatient bed days, the average prediction increased with more bed days, peaking at 14-29 days for all models except MARS. For mean HbA_{1c} values, the average prediction increased with higher HbA_{1c} levels, although a U-shaped relationship was observed for MARS, boosted trees, and MLP, with the lowest average predictions around HbA_{1c} levels of 6%-7%. Regarding diabetes medication categories, patients on insulin only and those on both oral diabetic medications and insulin had higher average predictions than those on oral medications only or no medications. PDPs for selected models showed that more advanced CKD stages (CKD stage 4 and stage 5) had higher average predictions. In most models, patients with ischemic heart disease or peripheral artery disease also had higher average predictions.

Regarding the prediction of ED visits ≥ 3 and ≥ 5 times (Figures S8-S9 in [Multimedia Appendix 1](#)), the selected models showed similar observations for age, present year ED visits, mean HbA_{1c}, diabetes medication categories, ischemic heart disease, and peripheral artery disease. It is noteworthy that present-year inpatient bed days did not significantly affect the predicted probability of these outcomes. For the prediction of ED visits ≥ 10 (Figure S10 in [Multimedia Appendix 1](#)), the PDPs aligned with the findings from both feature importance methods where the number of present year ED visits had the largest influence on average predictions, while other variables had smaller influence on average predictions.

Discussion

Principal Findings

In this study, we developed machine learning models to predict future inpatient and ED utilization by using sociodemographic characteristics, health indicators, diabetes-related complications, and prior utilization data from a chronic disease registry. We detailed a systematic approach to building, validating, and testing the models. Using this approach, we noted that imbalanced data distribution significantly affected model performance, often resulting in low sensitivity despite acceptable AUC values. This finding highlights the importance of considering multiple metrics, including AUC, sensitivity (recall), and PPV (precision), during model selection. We found that improved model performance can be achieved by addressing imbalanced data distribution through oversampling. We observed that random oversampling resulted in better model performance than SMOTE. Among the models trained with random oversampling, logistic regression, MARS, boosted trees, and MLP models had the best performance. Additionally, explainability analyses provided insights into how the best performing models made predictions and showed that their learned patterns were consistent with real-world knowledge, thereby supporting the validity of the models.

Predicting Future Inpatient Bed Days and ED Visits

In our study, we used inpatient bed days and ED visits within a calendar year as service level indicators of high health care utilization. Service level utilization is important because our prior research demonstrated a rising trend in diabetes-related complications [12] and our country is experiencing persistent bed shortages and crowded EDs [26]. In this context, service level utilization indicators are useful to inform health intervention programs to ease the bed crunch and overcrowded EDs. First, patients predicted to have very high level of health care utilization (ie, inpatient bed days ≥ 30 or ED visits ≥ 10) could be candidates for intensive case management to identify potential causes for prolonged admissions or frequent ED visits. Second, patients predicted to have moderately high level of health care utilization (ie, inpatient bed days ≥ 14 and < 30 and ED visits ≥ 5 and < 10) could be candidates for multidisciplinary (medical and social) diabetes care programs to reduce future utilization. Finally, patients with mildly elevated health care utilization (ie, inpatient bed days ≥ 7 and < 14 and ED visits ≥ 3 and < 5) could be candidates for novel care models that leverage technological solutions such as the Mobile Inpatient Care at Home [27].

Addressing Imbalanced Data Distribution by Using Data Sampling Approaches

Our study highlights the importance of addressing imbalanced data when developing machine learning models for health care applications. We observed that class imbalance can lead to acceptable AUC but low sensitivity—a phenomenon also noted in related literature [28]. Our study evaluates 2 different oversampling techniques: random oversampling and SMOTE. When comparing random oversampling with the 2 iterations of SMOTE, we found that random oversampling performed better than SMOTE ($k=3$), which in turn performed better than

SMOTE ($k=5$). This could suggest that predictive models perform better when the synthetic minority class used for training is similar to the actual training data. Random oversampling duplicates existing instances, whereas SMOTE ($k=3$) and SMOTE ($k=5$) create a new synthetic minority class by interpolating between 3 and 5 closely related minority class instances, respectively. It is recognized that with oversampling techniques, models may overfit and perform poorly in other data sets [14]. To investigate this, we tested our models on 2 additional test data sets (years 2020-2021 and 2021-2022) and found no degradation in model performance. Our conclusions were that because the training data were sufficiently large, it had good quality and variety to avoid overfitting.

Machine Learning Model Performance

Among the 7 machine learning models we tested, logistic regression, MARS, boosted trees, and MLP showed promising performance in predicting LOS across all 3 thresholds. For predicting ≥ 5 and ≥ 10 ED visits in the subsequent year, MARS and logistic regression outperformed the other models. Interestingly, logistic regression was found to be as effective as or even superior to other machine learning models in predicting health care utilization. These findings are noteworthy because while some studies have shown machine learning models to outperform traditional regression models in predicting health care utilization [3,28], others have found that machine learning models offered only limited improvement over traditional logistic regression [29]. When analyzing the model-specific variable importance scores and permutation feature importance plots for the selected models, we observed differences in the rankings of the important variables between models. However, the top 5 variables were generally consistent across selected models (Table S7 and Figures S3-S4 in [Multimedia Appendix 1](#)). In predicting inpatient LOS at all 3 thresholds, age, number of ED visits (present year), CKD stage, and inpatient bed days were the top 5 most important variables across all models. For predicting ED visits at all thresholds, the number of ED visits (present year), CKD stage, age, and mean HbA_{1c} values were the top 5 variables.

Additionally, explainability analyses using PDPs confirm what is known about high health care utilizers. Age, prior utilization in terms of ED visits and inpatient stays, and the presence of comorbidities and diabetes-related complications such as advanced stages of CKD, ischemic heart disease, and peripheral artery disease are associated with increased health care utilization. These findings suggest that current utilization is an important predictor of future utilization—a conclusion supported by similar studies [4,28]. Additionally, kidney disease has emerged as a significant predictor for future health care utilization in our cohort of patients with diabetes, as demonstrated in a recent study involving patients from the same population [30].

Interestingly, the U-shaped relationship between average prediction and HbA_{1c} values seen in many of the PDPs suggest that tight glycemic control (HbA_{1c}<6%) and relaxed glycemic control (HbA_{1c}≥8%) are associated with increased health care utilization. This is an interesting finding because we documented a similar U-shaped relationship previously between HbA_{1c} and

incidence of diabetes mellitus–related complications in the SDR [23]. Incident complications are expected to result in ED visits or admissions. Taken together, our explainability analyses suggest that the learned patterns are consistent with real-world knowledge and therefore lend support to the validity of the model.

Study Strengths, Limitations, and Future Research

Our study's strengths include the use of a large multiethnic cohort and easily obtainable predictors with minimal missing data. By utilizing different thresholds of inpatient bed days and ED visits as model outcomes, our approach allows policy makers and program planners to target interventions based on the predicted need. Other practitioners intending to build predictive models for population health programs could consider a similar systematic approach to building, validating, testing, and understanding the models. Through this approach, we were able to mitigate the problems associated with class imbalance by exploring the outcomes of the 2 data sampling methods. We also validated the models across different time frames and demonstrated their validity on unseen data. Finally, our explainability analyses provided reassurance that the models were making prediction based on learned patterns consistent with real-world knowledge. However, the absence of financial data and the nonexploration of other class imbalance methods such as feature selection are key limitations that could be addressed in future studies. Our test data sets spanned the COVID-19 pandemic, a period that may have affected health-seeking behavior and health care utilization. However, the consistency of our results with those from the validation data set, which was less affected by the pandemic, suggests that these potential anomalies did not significantly impact our findings. Another potential limitation is the exclusion of patients with missing data. In the context of this study, these patients are likely to be those who are well and had minimal interaction with the health system within that year. Given the large size of the data set for this study and the significant class imbalance for patients without any of the outcomes, it is likely that excluding patients due to missing data had minimal impact on model performance.

Although our study shortlisted 4 machine learning models with similar performance across different outcomes, it remains unclear which model is the most optimal. Beyond the performance variables, we considered the confusion matrix for each of the models and observed that these models describe alternative courses of action, each with a different cost and benefit attached; we will explore this in future research. Although we have described how the results from the models can be used in practice, we acknowledge the need for a more integrated approach to model selection and decision-making criteria. In this regard, we are currently exploring additional methods to address this, specifically focusing on how to combine the outputs of the binary classification models into a single more comprehensive multiclass prediction model. To achieve this, we are investigating the use of hierarchical decision models and ensemble model approaches. These methods would allow us to integrate the predictions from individual binary models into a unified multiclass model, making it more applicable in real-world scenarios. However, these additional methods and

their applications will be detailed in a follow-up study. Relatedly, the models that we developed are predictive and they are unable to provide prescriptive insights. Additional tools will be needed to be developed to profile patients and identify the most appropriate interventions for them. Finally, since our study uses data from a public regional health database in Singapore, the findings may not be generalizable to other contexts.

Conclusion

We were able to apply common machine learning algorithms to predict future health care utilization by using inpatient bed days and ED utilization as the predicted outcomes. These predictive models will be useful to policy makers and program planners as they develop population health initiatives to improve care for patients with diabetes and manage acute health care utilization.

Data Availability

Data cannot be shared publicly because of Singapore Health Services Cluster Data Policy on data sharing restrictions. Data are available from the Singapore Health Services Diabetes Registry Committee for researchers who meet the criteria for access to confidential data. The criteria include institutional review board approval, data use approval, and a research collaboration agreement. The point of contact for the Singapore Health Services Diabetes Registry Disease Registry Committee is Ms Lee Thong Shuen (email: lee.thong.shuen@singhealth.com.sg).

Authors' Contributions

JKT contributed to study design, statistical analyses, model development, data interpretation, data visualization, manuscript drafting, and review. LQ contributed to study design, model development, data interpretation, and review. NNMS contributed to data collection and manuscript review. JHT contributed to study design, model development, data interpretation, data visualization, and review. Goh SY, JT, and YMB contributed to study design, data interpretation, and manuscript review.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Supplementary data.

[[DOCX File, 5972 KB](#) - [ai_v3i1e58463_app1.docx](#)]

Multimedia Appendix 2

High-resolution images of Figures 1-4.

[[PDF File \(Adobe PDF File\), 6818 KB](#) - [ai_v3i1e58463_app2.pdf](#)]

References

1. Singapore healthcare. International Trade Association Market Intelligence. URL: <https://www.trade.gov/market-intelligence/singapore-healthcare> [accessed 2023-10-23]
2. White Paper on Healthier SG. Ministry of Health Singapore. URL: <https://www.healthiersg.gov.sg/resources/white-paper/> [accessed 2023-10-23]
3. Osawa I, Goto T, Yamamoto Y, Tsugawa Y. Machine-learning-based prediction models for high-need high-cost patients using nationwide clinical and claims data. NPJ Digit Med 2020 Nov 11;3(1):148 [FREE Full text] [doi: [10.1038/s41746-020-00354-8](https://doi.org/10.1038/s41746-020-00354-8)] [Medline: [33299137](https://pubmed.ncbi.nlm.nih.gov/33299137/)]
4. Langenberger B, Schulte T, Groene O. The application of machine learning to predict high-cost patients: A performance-comparison of different models using healthcare claims data. PLoS One 2023;18(1):e0279540 [FREE Full text] [doi: [10.1371/journal.pone.0279540](https://doi.org/10.1371/journal.pone.0279540)] [Medline: [36652450](https://pubmed.ncbi.nlm.nih.gov/36652450/)]
5. Das L, Abramson E, Kaushal R. High-need, high-cost patients offer solutions for improving their care and reducing costs. NEJM Catal. 2019 Feb 05. URL: <https://catalyst.nejm.org/doi/full/10.1056/CAT.19.0015> [accessed 2023-10-23]
6. Madvig P, Pearl R. Managing the most expensive patients. Harvard Business Review. URL: <https://hbr.org/2020/01/managing-the-most-expensive-patients> [accessed 2023-10-23]
7. Wodchis WP, Austin PC, Henry DA. A 3-year study of high-cost users of health care. CMAJ 2016 Feb 16;188(3):182-188 [FREE Full text] [doi: [10.1503/cmaj.150064](https://doi.org/10.1503/cmaj.150064)] [Medline: [26755672](https://pubmed.ncbi.nlm.nih.gov/26755672/)]
8. Ng SH, Rahman N, Ang IYH, Sridharan S, Ramachandran S, Wang DD, et al. Characterization of high healthcare utilizer groups using administrative data from an electronic medical record database. BMC Health Serv Res 2019 Jul 05;19(1):452 [FREE Full text] [doi: [10.1186/s12913-019-4239-2](https://doi.org/10.1186/s12913-019-4239-2)] [Medline: [31277649](https://pubmed.ncbi.nlm.nih.gov/31277649/)]
9. de Ruijter UW, Kaplan ZLR, Bramer WM, Eijkenaar F, Nieboer D, van der Heide A, et al. Prediction models for future high-need high-cost healthcare use: a systematic review. J Gen Intern Med 2022 May;37(7):1763-1770 [FREE Full text] [doi: [10.1007/s11606-021-07333-z](https://doi.org/10.1007/s11606-021-07333-z)] [Medline: [35018571](https://pubmed.ncbi.nlm.nih.gov/35018571/)]

10. Lim DYZ, Chia SY, Abdul Kadir H, Mohamed Salim NN, Bee YM. Establishment of the SingHealth Diabetes Registry. *CLEP* 2021 Mar;Volume 13:215-223. [doi: [10.2147/clep.s300663](https://doi.org/10.2147/clep.s300663)]
11. Shannon B, Pang R, Jepson M, Williams C, Andrew N, Smith K, et al. What is the prevalence of frequent attendance to emergency departments and what is the impact on emergency department utilisation? A systematic review and meta-analysis. *Intern Emerg Med* 2020 Oct;15(7):1303-1316. [doi: [10.1007/s11739-020-02403-2](https://doi.org/10.1007/s11739-020-02403-2)] [Medline: [32557095](https://pubmed.ncbi.nlm.nih.gov/32557095/)]
12. Tan JK, Salim NNM, Lim GH, Chia SY, Thumboo J, Bee YM. Trends in diabetes-related complications in Singapore, 2013-2020: A registry-based study. *PLoS One* 2022;17(10):e0275920 [FREE Full text] [doi: [10.1371/journal.pone.0275920](https://doi.org/10.1371/journal.pone.0275920)] [Medline: [36219616](https://pubmed.ncbi.nlm.nih.gov/36219616/)]
13. Type 2 diabetes mellitus - personalising management with non-insulin medications. Agency For Care Effectiveness. 2023 May 17. URL: [https://www.ace-hta.gov.sg/healthcare-professionals/ace-clinical-guidances-\(acgs\)/details/t2dm-personalising-medications](https://www.ace-hta.gov.sg/healthcare-professionals/ace-clinical-guidances-(acgs)/details/t2dm-personalising-medications) [accessed 2023-10-23]
14. Leevy JL, Khoshgofaar TM, Bauder RA, Seliya N. A survey on addressing high-class imbalance in big data. *J Big Data* 2018 Nov 1;5(1):1-30. [doi: [10.1186/s40537-018-0151-6](https://doi.org/10.1186/s40537-018-0151-6)]
15. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: Synthetic Minority Oversampling Technique. *Journal of Artificial Intelligence Research* 2002 Jun 01;16:321-357 [FREE Full text] [doi: [10.1613/jair.953](https://doi.org/10.1613/jair.953)]
16. Emil H. themis: extra recipes steps for dealing with unbalanced data. The Comprehensive R Archive Network. 2023 Aug 14. URL: <https://cran.r-project.org/web/packages/themis/index.html> [accessed 2023-10-23]
17. Fernández A, del Río S, Chawla NV, Herrera F. An insight into imbalanced big data classification: outcomes and challenges. *Complex Intell Syst* 2017 Mar 1;3(2):105-120. [doi: [10.1007/s40747-017-0037-9](https://doi.org/10.1007/s40747-017-0037-9)]
18. Max K, Jed W, Steve W, Andre W, Chris K. caret: classification and regression training. CRAN: Package caret. 2023 Mar 21. URL: <https://cran.r-project.org/web/packages/caret/index.html> [accessed 2023-10-23]
19. Ozenne B, Subtil F, Maucort-Boulch D. The precision-recall curve overcame the optimism of the receiver operating characteristic curve in rare diseases. *J Clin Epidemiol* 2015 Aug;68(8):855-859. [doi: [10.1016/j.jclinepi.2015.02.010](https://doi.org/10.1016/j.jclinepi.2015.02.010)] [Medline: [25881487](https://pubmed.ncbi.nlm.nih.gov/25881487/)]
20. Max K, Hadley W. Tidymodels: a collection of packages for modeling and machine learning using tidyverse principles. CRAN.r-Project. URL: <https://cran.r-project.org/web/packages/tidymodels/citation.html> [accessed 2023-10-23]
21. Greenwell B, Boehmke B. Variable importance plots—an introduction to the vip package. CRAN: Package vip. URL: <https://cran.r-project.org/web/packages/vip/index.html> [accessed 2023-10-23]
22. Biecek P, Maksymiuk S, Baniecki H. DALEX: moDel Agnostic Language for Exploration and eXplanation. CRAN: Package DALEX. URL: <https://cran.r-project.org/web/packages/DALEX/index.html> [accessed 2024-06-10]
23. Tan JK, Lim GH, Mohamed Salim NN, Chia SY, Thumboo J, Bee YM. Associations between mean HbA1c, HbA1c variability, and both mortality and macrovascular complications in patients with diabetes mellitus: a registry-based cohort study. *Clin Epidemiol* 2023;15:137-149 [FREE Full text] [doi: [10.2147/CLEP.S391749](https://doi.org/10.2147/CLEP.S391749)] [Medline: [36721457](https://pubmed.ncbi.nlm.nih.gov/36721457/)]
24. Klement W, El Emam K. Consolidated reporting guidelines for prognostic and diagnostic machine learning modeling studies: development and validation. *J Med Internet Res* 2023 Aug 31;25:e48763 [FREE Full text] [doi: [10.2196/48763](https://doi.org/10.2196/48763)] [Medline: [37651179](https://pubmed.ncbi.nlm.nih.gov/37651179/)]
25. Department of Statistics Singapore. Singapore residents by age group, ethnic group and sex, end June. SingStat Table Builder. URL: <https://tablebuilder.singstat.gov.sg/table/TS/M810011> [accessed 2023-10-23]
26. Update on efforts to alleviate shortage of hospital beds and nursing staff in public hospitals. Ministry of Health Singapore. URL: <https://www.moh.gov.sg/news-highlights/details/update-on-efforts-to-alleviate-shortage-of-hospital-beds-and-nursing-staff-in-public-hospitals> [accessed 2024-01-16]
27. Mobile inpatient care @ home sandbox to expand to more public hospitals to cover more medical conditions - MOH office for healthcare transformation. MOHT. URL: <https://moht.com.sg/mobile-inpatient-care-home-sandbox-to-expand-to-more-public-hospitals-to-cover-more-medical-conditions/> [accessed 2023-10-23]
28. Nghiem N, Atkinson J, Nguyen BP, Tran-Duy A, Wilson N. Predicting high health-cost users among people with cardiovascular disease using machine learning and nationwide linked social administrative datasets. *Health Econ Rev* 2023 Feb 04;13(1):9 [FREE Full text] [doi: [10.1186/s13561-023-00422-1](https://doi.org/10.1186/s13561-023-00422-1)] [Medline: [36738348](https://pubmed.ncbi.nlm.nih.gov/36738348/)]
29. Desai RJ, Wang SV, Vaduganathan M, Evers T, Schneeweiss S. Comparison of machine learning methods with traditional models for use of administrative claims with electronic medical records to predict heart failure outcomes. *JAMA Netw Open* 2020 Jan 03;3(1):e1918962 [FREE Full text] [doi: [10.1001/jamanetworkopen.2019.18962](https://doi.org/10.1001/jamanetworkopen.2019.18962)] [Medline: [31922560](https://pubmed.ncbi.nlm.nih.gov/31922560/)]
30. Tan JK, Kadir HA, Lim GH, Thumboo J, Bee YM, Lim CC. Trends in fluid overload-related hospitalisations among patients with diabetes mellitus The impact of chronic kidney disease. *Ann Acad Med Singap* 2024 Jul 30;53(7):435-445 [FREE Full text] [doi: [10.47102/annals-acadmedsg.2024136](https://doi.org/10.47102/annals-acadmedsg.2024136)] [Medline: [39132960](https://pubmed.ncbi.nlm.nih.gov/39132960/)]

Abbreviations

- AUC:** area under the receiver operating characteristic curve
CKD: chronic kidney disease

ED: emergency department
HbA_{1c}: hemoglobin A_{1c}
HNHC: high-need, high-cost
LOS: length of stay
MARS: multivariate adaptive regression splines
MLP: multilayer perceptron
PDP: partial dependence plot
PPV: positive predictive value
SDR: SingHealth Diabetes Registry
SingHealth: Singapore Health Services
SMOTE-NC: synthetic minority oversampling technique-nominal continuous

Edited by K El Emam, B Malin; submitted 16.03.24; peer-reviewed by Y Wang, S Mao, U Sinha; comments to author 06.06.24; revised version received 24.07.24; accepted 24.08.24; published 17.10.24.

Please cite as:

Tan JK, Quan L, Salim NNM, Tan JH, Goh SY, Thumboo J, Bee YM

Machine Learning–Based Prediction for High Health Care Utilizers by Using a Multi-Institutional Diabetes Registry: Model Training and Evaluation

JMIR AI 2024;3:e58463

URL: <https://ai.jmir.org/2024/1/e58463>

doi: [10.2196/58463](https://doi.org/10.2196/58463)

PMID:

©Joshua Kuan Tan, Le Quan, Nur Nasyitah Mohamed Salim, Jen Hong Tan, Su-Yen Goh, Julian Thumboo, Yong Mong Bee. Originally published in JMIR AI (<https://ai.jmir.org>), 17.10.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR AI, is properly cited. The complete bibliographic information, a link to the original publication on <https://www.ai.jmir.org/>, as well as this copyright and license information must be included.

Research Letter

Can Large Language Models Replace Therapists? Evaluating Performance at Simple Cognitive Behavioral Therapy Tasks

Nathan Hodson¹, BMBS, MPH, MRCPsych; Simon Williamson¹, MBBS, MPhil

Warwick Medical School, University of Warwick, Coventry, United Kingdom

Corresponding Author:

Nathan Hodson, BMBS, MPH, MRCPsych

Warwick Medical School

University of Warwick

Warwick Medical School

Gibbett Hill Road

Coventry, CV4 7AL

United Kingdom

Phone: 44 02476574880

Email: nathan.hodson@warwick.ac.uk

Abstract

The advent of large language models (LLMs) such as ChatGPT has potential implications for psychological therapies such as cognitive behavioral therapy (CBT). We systematically investigated whether LLMs could recognize an unhelpful thought, examine its validity, and reframe it to a more helpful one. LLMs currently have the potential to offer reasonable suggestions for the identification and reframing of unhelpful thoughts but should not be relied on to lead CBT delivery.

(*JMIR AI* 2024;3:e52500) doi:[10.2196/52500](https://doi.org/10.2196/52500)

KEYWORDS

mental health; psychotherapy; digital therapy; CBT; ChatGPT; cognitive behavioral therapy; cognitive behavioural therapy; LLM; LLMs; language model; language models; NLP; natural language processing; artificial intelligence; performance; chatbot; chatbots; conversational agent; conversational agents

Introduction

Large language models (LLMs) represent a significant advance in the field of artificial intelligence (AI) and herald a transformational change in the role of computers both personally and professionally. LLMs, such as OpenAI's ChatGPT and Google's Bard (later rebranded as Gemini), represent a new form of generative AI. They have linguistic capabilities comparable to humans, and they demonstrate performance similar to specialized models for sentiment analysis and affective computing [1]. Psychiatry and psychology, and talking therapy, in particular, is a field with significant potential impact of LLMs. Demand for therapists greatly outweighs supply, making the question of how new technologies could relieve pressure on mental health systems a pertinent one. Here we report an evaluation of whether existing LLMs can contribute to the delivery of cognitive behavioral therapy (CBT), and their limitations.

CBT is a first-line treatment for common mental health disorders, including anxiety and depression. It involves understanding cognitive biases and challenging those thoughts.

Where other modes of psychotherapy rely on the therapist's individualized interpretation, CBT emphasizes systematic changes in thinking and behavior.

Self-guided, web-based CBT has emerged as a response to the shortage of CBT therapists, and it is increasingly recommended as an accessible alternative [2]. These programs reduce the input of the human therapist to a brief phone call, with patients assigned web-based modules to complete. Although the approach is cost-effective and scalable, it risks making the content of web-based CBT less personalized. Since LLMs can flexibly respond to personal circumstances, they may be well-suited to addressing this.

AI has previously been used to augment CBT by performing peripheral tasks. In a study of chronic pain, AI was used to select the appropriate CBT intervention for patients each week based on the previous week's progress [2]. The digital CBT company Wysa [3] uses AI to select appropriate therapist-authored responses. Mental Health America has built a website using AI to help people identify and reframe cognitive biases as an isolated exercise [4]. However, none of these applications have

harnessed the generative capacity of LLMs as therapeutic chatbots to aid patients in reframing unhelpful thoughts.

We aimed to understand whether AI could recognize an unhelpful thought, examine its validity, and reframe it to a more helpful one. This technique, often referred to as “catch it, check it, change it,” requires knowledge of cognitive biases, the linguistic ability to reframe them, and importantly, a degree of comprehension such that the reframing meaningfully addresses the bias [5]. If publicly available LLMs can support “Catch It, Check It, Change It,” then they may have a valuable role in increasing the effectiveness of digital CBT.

Methods

We explored whether OpenAI’s ChatGPT-4 and Google’s Bard could perform the 3 stages of the “catch it, check it, change it” technique (see Table 1). Two independent CBT therapists currently practising in the UK’s National Health Service aided in assessing the LLMs, rating whether they had completed the tasks satisfactorily. The therapists each wrote their own set of 10 thoughts, ensuring they received different replies from the LLMs. Both ChatGPT-4 and Bard responded to 20 tasks at each stage of the study. The sessions for each therapist occurred on June 2 and 14, 2023.

Table 1. Evaluating how large language models (LLMs) perform at the Catch It, Check It, Change It approach.

	CBT ^a skill	Input to LLM	Task for LLM	Criteria
Stage 1: “Catch it”	“Catch it” means patients can stop and notice that their thought may be distorted. Therapists must be able to illustrate different distortions.	Titles of 10 cognitive biases	Generate a two-sentence vignette for each bias.	Could CBT therapists work out which bias each vignette illustrated?
Stage 2: “Check it”	“Check it” means patients consider whether a thought is helpful, or whether it fits with a cognitive distortion. Therapists must be able to explain which distortion a thought fits into.	Therapist-written thoughts illustrating 10 cognitive distortions, each in the language of a patient. Each therapist produced an independent list of thoughts with no discussion.	Identify which cognitive bias each vignette represents.	Did LLMs identify the same biases?
Stage 3: “Change it”	“Change it” means patients can reframe their thoughts. Therapists should be able to suggest reframing of thoughts that patients may consider.	Therapist-written thoughts illustrating 10 cognitive biases as above	Reframe the thought to overcome the bias.	Did therapists think the new thought addressed the bias?

^aCBT: cognitive behavioral therapy.

Results

Table 2 shows LLM performance over the 3 tasks. Both models demonstrated varying levels of proficiency across tasks and therapists. Overall, ChatGPT-4 scored 44/60 and Bard scored 42/60. Both performed similarly at generating vignettes, which clearly illustrated a cognitive bias (Stage 1: ChatGPT 13/20, Bard 13/20), whereas ChatGPT-4 performed better at identifying cognitive biases (Stage 2: ChatGPT 15/20, Bard 10/20). The LLMs performed superiorly at reframing unhelpful thoughts, with Bard achieving a near-perfect score (Stage 3: ChatGPT 16/20, Bard 19/20).

Frequently, the LLMs were only marginally incorrect. Specifically, Bard often mentioned cognitive biases outside of the 10 provided, using alternative labels that nonetheless described the bias plausibly. This may reflect an inherent limitation of CBT terminology, rather than poor model performance. Indeed, this limitation appeared to extend to therapists, who only demonstrated moderate inter-rater reliability in labeling LLM-generated vignettes (Cohen $\kappa=0.44$). However, at stage 3, therapist 2 noted several instances where the LLM “missed the point” and, while technically improving the original thought, did not reframe it in a way that demonstrated understanding of the underlying cognitive bias. Prompts given to these LLMs and examples of errors noted in the outputs are presented in Multimedia Appendix 1.

Table 2. Number of tasks completed correctly at each stage.

Evaluation stage	Bard			ChatGPT-4		
	Therapist 1 (out of 10)	Therapist 2 (out of 10)	Total	Therapist 1 (out of 10)	Therapist 2 (out of 10)	Total
Stage 1: Catch it (How many LLM-generated vignettes were correctly identified by a therapist?)	7	6	13	8	5	13
Stage 2: Check it (How many therapist-generated vignettes were correctly identified by the LLM?)	7	3	10	7	8	15
Stage 3: Change it (How many LLM-reformulated vignettes were considered improvements by a therapist?)	10	9	19	10	6	16

Discussion

Our study findings suggest that LLMs should not yet be relied on to lead CBT delivery, although LLMs show clear potential as assistants capable of offering reasonable suggestions for the identification and reframing of unhelpful thoughts.

LLMs are far from replacing CBT therapists, but they perform well in some isolated tasks (eg, Bard for reframing), so it is worthwhile exploring limited yet innovative ways to use AI to improve patient experience and outcomes. We suggest CBT therapists equip patients with a working knowledge of cognitive biases, but therapists could also advise patients to consider using LLMs to gather suggestions on reframing unhelpful thoughts beyond sessions.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Prompts provided to the large language models Bard and ChatGPT-4 and examples of errors noted in the outputs.

[[DOCX File, 18 KB - ai_v3i1e52500_app1.docx](#)]

References

1. Amin MM, Cambria E, Schuller BW. Will affective computing emerge from foundation models and general artificial intelligence? A first evaluation of ChatGPT. *IEEE Intell Syst* 2023 Mar;38(2):15-23. [doi: [10.1109/mis.2023.3254179](https://doi.org/10.1109/mis.2023.3254179)]
2. Guided self-help digital cognitive behavioural therapy for children and young people with mild to moderate symptoms of anxiety or low mood: early value assessment. *Health Technology Evaluation*.: National Institute for Health and Care Excellence; 2023. URL: <https://www.nice.org.uk/guidance/hte3> [accessed 2024-07-22]
3. Wysa - everyday mental health. URL: <https://www.wysa.com/> [accessed 2024-07-22]
4. Overcoming negative thoughts. *Mental Health America*. 2023. URL: https://screening.mhanational.org/diy/overcoming-negative-thoughts/?layout=actions_neutral [accessed 2023-06-20]
5. Cognitive change exercise. *Think CBT*. 2022. URL: https://thinkcbt.com/images/CATCH_CHECK_CHANGE_EXERCISE.pdf [accessed 2023-06-20]

Edited by K El Emam, B Malin; submitted 06.09.23; peer-reviewed by R Yang, L Magoun, L Zhu; comments to author 24.10.23; revised version received 14.12.23; accepted 01.01.24; published 30.07.24.

Please cite as:

Hodson N, Williamson S

Can Large Language Models Replace Therapists? Evaluating Performance at Simple Cognitive Behavioral Therapy Tasks

JMIR AI 2024;3:e52500

URL: <https://ai.jmir.org/2024/1/e52500>

doi: [10.2196/52500](https://doi.org/10.2196/52500)

PMID: [39078696](https://pubmed.ncbi.nlm.nih.gov/39078696/)

©Nathan Hodson, Simon Williamson. Originally published in JMIR AI (<https://ai.jmir.org>), 30.07.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR AI, is properly cited. The complete bibliographic information, a link to the original publication on <https://www.ai.jmir.org/>, as well as this copyright and license information must be included.

Corrigenda and Addenda

Correction: Using Conversational AI to Facilitate Mental Health Assessments and Improve Clinical Efficiency Within Psychotherapy Services: Real-World Observational Study

Max Rollwage¹, BSc, MSc, MPhil, PhD; Johanna Habicht¹, MSci; Keno Juechems¹, BSc, MSc, PhD; Ben Carrington¹, BSc; Sruthi Viswanathan¹, BTech, MRes; Mona Stylianou², BA, PGDip; Tobias U Hauser^{1,3,4,5}, PhD; Ross Harper¹, MA, MRes, PhD

¹Limbic Limited, London, United Kingdom

²Everyturn Mental Health, Gosforth, United Kingdom

³Max Planck University College London Centre for Computational Psychiatry and Ageing Research, University College London, London, United Kingdom

⁴Department of Psychiatry and Psychotherapy, Medical School and University Hospital, Eberhard Karls University of Tübingen, Tübingen, Germany

⁵German Center for Mental Health (DZPG), Tübingen, Germany

Corresponding Author:

Max Rollwage, BSc, MSc, MPhil, PhD

Limbic Limited

Kemp House

160 City Road

London,

United Kingdom

Phone: 44 07491263783

Email: max@limbic.ai

Related Article:

Correction of: <https://ai.jmir.org/2023/1/e44358>

(*JMIR AI* 2024;3:e57869) doi:[10.2196/57869](https://doi.org/10.2196/57869)

In “Using Conversational AI to Facilitate Mental Health Assessments and Improve Clinical Efficiency Within Psychotherapy Services: Real-World Observational Study” (*JMIR AI* 2023;2:e44358) the authors noted one error.

One author, Sruthi Viswanathan, was inadvertently omitted from the authorship list in the original publication of the paper. Sruthi Viswanathan has now been added to the authorship of the published paper as the fifth author, with the degrees "BTech, MRes" and the following affiliation:

Limbic Limited, London, United Kingdom

In accordance, the Conflict of Interest statement has also been updated to include this author. The originally published statement appeared as follows:

MR, KJ, JH, BC, and RH are employed by Limbic Limited and hold shares in the company. TUH works as a paid consultant for Limbic Limited and holds shares in the company.

This statement has been corrected to:

MR, KJ, JH, BC, SV and RH are employed by Limbic Limited and hold shares in the company. TUH works as a paid consultant for Limbic Limited and holds shares in the company.

The correction will appear in the online version of the paper on the JMIR Publications website on March 12, 2024, together with the publication of this correction notice. Because this was made after submission to PubMed, PubMed Central, and other full-text repositories, the corrected article has also been resubmitted to those repositories.

Submitted 29.02.24; this is a non-peer-reviewed article; accepted 01.03.24; published 12.03.24.

Please cite as:

Rollwage M, Habicht J, Juechems K, Carrington B, Viswanathan S, Stylianou M, Hauser TU, Harper R

Correction: Using Conversational AI to Facilitate Mental Health Assessments and Improve Clinical Efficiency Within Psychotherapy Services: Real-World Observational Study

JMIR AI 2024;3:e57869

URL: <https://ai.jmir.org/2024/1/e57869>

doi: [10.2196/57869](https://doi.org/10.2196/57869)

PMID:

©Max Rollwage, Johanna Habicht, Keno Juechems, Ben Carrington, Sruthi Viswanathan, Mona Stylianou, Tobias U Hauser, Ross Harper. Originally published in JMIR AI (<https://ai.jmir.org>), 12.03.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR AI, is properly cited. The complete bibliographic information, a link to the original publication on <https://www.ai.jmir.org/>, as well as this copyright and license information must be included.

Corrigenda and Addenda

Correction: Feasibility of Multimodal Artificial Intelligence Using GPT-4 Vision for the Classification of Middle Ear Disease: Qualitative Study and Validation

Masao Noda^{1*}, MBA, MD, PhD; Hidekane Yoshimura^{2*}, MD, PhD; Takuya Okubo², MD; Ryota Koshu¹, MD; Yuki Uchiyama², MD; Akihiro Nomura³, MD, PhD; Makoto Ito¹, MD, PhD; Yutaka Takumi², MD, PhD

¹Department of Otolaryngology, Head and Neck Surgery, Jichi Medical University, Shimotsuke, Japan

²Department of Otolaryngology - Head and Neck Surgery, Shinshu University, Matsumoto, Japan

³College of Transdisciplinary Sciences for Innovation, Kanazawa University, Kanazawa, Japan

*these authors contributed equally

Corresponding Author:

Masao Noda, MBA, MD, PhD

Department of Otolaryngology, Head and Neck Surgery

Jichi Medical University

3311-1 Yakushiji

Shimotsuke, 329-0498

Japan

Phone: 81 285442111

Email: doforanabdosuc@gmail.com

Related Article:

Correction of: <https://ai.jmir.org/2024/1/e58342>

(*JMIR AI 2024;3:e62990*) doi:[10.2196/62990](https://doi.org/10.2196/62990)

In “Feasibility of Multimodal Artificial Intelligence Using GPT-4 Vision for the Classification of Middle Ear Disease: Qualitative Study and Validation” (JMIR AI 2024;3:e58342) the authors noted one error.

In the original author, no equal contributors were specified. This has been changed as follows:

Masao Noda, Hidekane Yoshimura**

**These authors contributed equally*

The correction will appear in the online version of the paper on the JMIR Publications website on July 9, 2024, together with the publication of this correction notice. Because this was made after submission to PubMed, PubMed Central, and other full-text repositories, the corrected article has also been resubmitted to those repositories.

Submitted 06.06.24; this is a non-peer-reviewed article; accepted 06.06.24; published 09.07.24.

Please cite as:

Noda M, Yoshimura H, Okubo T, Koshu R, Uchiyama Y, Nomura A, Ito M, Takumi Y

Correction: Feasibility of Multimodal Artificial Intelligence Using GPT-4 Vision for the Classification of Middle Ear Disease: Qualitative Study and Validation

JMIR AI 2024;3:e62990

URL: <https://ai.jmir.org/2024/1/e62990>

doi: [10.2196/62990](https://doi.org/10.2196/62990)

PMID:

©Masao Noda, Hidekane Yoshimura, Takuya Okubo, Ryota Koshu, Yuki Uchiyama, Akihiro Nomura, Makoto Ito, Yutaka Takumi. Originally published in JMIR AI (<https://ai.jmir.org>), 09.07.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted

use, distribution, and reproduction in any medium, provided the original work, first published in JMIR AI, is properly cited. The complete bibliographic information, a link to the original publication on <https://www.ai.jmir.org/>, as well as this copyright and license information must be included.

Original Paper

Predictive Performance of Machine Learning–Based Models for Poststroke Clinical Outcomes in Comparison With Conventional Prognostic Scores: Multicenter, Hospital-Based Observational Study

Fumi Irie^{1,2*}, MD, PhD; Koutarou Matsumoto^{3*}, MPH, PhD; Ryu Matsuo^{1,2}, MD, PhD; Yasunobu Nohara⁴, PhD; Yoshinobu Wakisaka², MD, PhD; Tetsuro Ago^{2,5}, MD, PhD; Naoki Nakashima⁶, MD, PhD; Takanari Kitazono^{2,5}, MD, PhD; Masahiro Kamouchi^{1,5}, MD, PhD[‡]

¹Department of Health Care Administration and Management, Graduate School of Medical Sciences, Kyushu University, Fukuoka, Japan

²Department of Medicine and Clinical Science, Graduate School of Medical Sciences, Kyushu University, Fukuoka, Japan

³Biostatistics Center, Graduate School of Medicine, Kurume University, Kurume, Japan

⁴Big Data Science and Technology, Faculty of Advanced Science and Technology, Kumamoto University, Kumamoto, Japan

⁵Center for Cohort Studies, Graduate School of Medical Sciences, Kyushu University, Fukuoka, Japan

⁶Medical Information Center, Kyushu University Hospital, Fukuoka, Japan

[‡]Fukuoka Stroke Registry Investigators

*these authors contributed equally

Corresponding Author:

Masahiro Kamouchi, MD, PhD

Department of Health Care Administration and Management

Graduate School of Medical Sciences

Kyushu University

3-1-1 Maidashi

Higashi-ku

Fukuoka, 812-8582

Japan

Phone: 81 92 642 6960

Email: kamouchi.masahiro.736@m.kyushu-u.ac.jp

Abstract

Background: Although machine learning is a promising tool for making prognoses, the performance of machine learning in predicting outcomes after stroke remains to be examined.

Objective: This study aims to examine how much data-driven models with machine learning improve predictive performance for poststroke outcomes compared with conventional stroke prognostic scores and to elucidate how explanatory variables in machine learning–based models differ from the items of the stroke prognostic scores.

Methods: We used data from 10,513 patients who were registered in a multicenter prospective stroke registry in Japan between 2007 and 2017. The outcomes were poor functional outcome (modified Rankin Scale score >2) and death at 3 months after stroke. Machine learning–based models were developed using all variables with regularization methods, random forests, or boosted trees. We selected 3 stroke prognostic scores, namely, ASTRAL (Acute Stroke Registry and Analysis of Lausanne), PLAN (preadmission comorbidities, level of consciousness, age, neurologic deficit), and iScore (Ischemic Stroke Predictive Risk Score) for comparison. Item-based regression models were developed using the items of these 3 scores. The model performance was assessed in terms of discrimination and calibration. To compare the predictive performance of the data-driven model with that of the item-based model, we performed internal validation after random splits of identical populations into 80% of patients as a training set and 20% of patients as a test set; the models were developed in the training set and were validated in the test set. We evaluated the contribution of each variable to the models and compared the predictors used in the machine learning–based models with the items of the stroke prognostic scores.

Results: The mean age of the study patients was 73.0 (SD 12.5) years, and 59.1% (6209/10,513) of them were men. The area under the receiver operating characteristic curves and the area under the precision-recall curves for predicting poststroke outcomes were higher for machine learning–based models than for item-based models in identical populations after random splits. Machine learning–based models also performed better than item-based models in terms of the Brier score. Machine learning–based models used different explanatory variables, such as laboratory data, from the items of the conventional stroke prognostic scores. Including these data in the machine learning–based models as explanatory variables improved performance in predicting outcomes after stroke, especially poststroke death.

Conclusions: Machine learning–based models performed better in predicting poststroke outcomes than regression models using the items of conventional stroke prognostic scores, although they required additional variables, such as laboratory data, to attain improved performance. Further studies are warranted to validate the usefulness of machine learning in clinical settings.

(JMIR AI 2024;3:e46840) doi:[10.2196/46840](https://doi.org/10.2196/46840)

KEYWORDS

brain infarction; outcome; prediction; machine learning; prognostic score

Introduction

Background

Despite receiving the best available treatment, patients who have had a stroke may still experience disability or, in some cases, even face the risk of death [1,2]. Stroke clinicians try to predict patients' outcomes as accurately as possible because accurate prognoses are a prerequisite for therapeutic decisions. Various stroke prognostic scores have been developed to support clinicians in predicting poststroke outcomes [3-8]. Nevertheless, prognostic scores have some disadvantages: generally, they limit the number of variables for ease of use at the bedside, and their validity needs to be reappraised over time, as the scoring criteria may become outdated with rapid progress in stroke care [9].

Meanwhile, recent advances in information technology have enabled the collection of a large amount of health information on individual patients [10,11]. Machine learning is considered a promising tool for improving the prediction accuracy of clinical outcomes for individual patients with stroke because of the ability of machine learning to deal with large and complex data [12-24].

However, several papers questioning the incremental value of machine learning have recently been published [25-27]. One study reported that machine learning algorithms did not perform better than traditional regression models for making prognoses in traumatic brain injury and recommended replicating studies in fields other than traumatic brain injury to ensure the generalizability of the findings [26]. Hitherto, few studies have directly compared the performance of data-driven models developed using machine learning methods and regression models based on conventional stroke prognostic scores in the field of outcome prediction after ischemic stroke [19,20,23]. In addition, calibration has not been adequately addressed in previous studies, and model performance has primarily been evaluated based on its discriminative ability [18-20].

Objectives

In this study, we aimed to examine whether machine learning can improve the predictive performance for poststroke outcomes beyond preexisting stroke prognostic scores. We also sought to

elucidate the pattern of variables selected by machine learning algorithms to predict poststroke clinical outcomes. To this end, we analyzed the data of patients with acute ischemic stroke enrolled in a multicenter, hospital-based, prospective registry of stroke in Japan. We used 3 stroke prognostic scores, namely, Acute Stroke Registry and Analysis of Lausanne (ASTRAL) score [6], preadmission comorbidities, level of consciousness, age, and neurologic deficit (PLAN) score [7], and Ischemic Stroke Predictive Risk Score (iScore) [4,5], to create item-based regression models. We then compared the predictive performance of data-driven models developed using machine learning algorithms with that of item-based models in identical study populations. We also examined the explanatory variables used in data-driven models and compared them with the items of the conventional prognostic scores.

Methods

Ethical Considerations

The study protocol was approved by the institutional review boards of all hospitals (Kyushu University Institutional Review Board for Clinical Research: 22086-01; Kyushu Medical Center Institutional Review Board: R06-03; Clinical Research Review Board of Fukuokahigashi Medical Center: 29-C-38; Fukuoka Red Cross Hospital Institutional Review Board: 629; St Mary's Hospital Research Ethics Review Committee: S13-0110; Steel Memorial Yawata Hospital Ethics Committee: 06-04-13; and Kyushu Rosai Hospital Institutional Review Board: 21-8). Written informed consent was obtained from all patients or their family members.

Data Source

We used data from the Fukuoka Stroke Registry (FSR), a multicenter, hospital-based, prospective registry of patients with acute stroke. FSR enrolled patients with stroke hospitalized in 7 participating hospitals in Fukuoka, Japan, within 7 days of onset (University Hospital Medical Information Network Clinical Trial Registry: UMIN000000800). Details of the registry have been previously published [28,29]. In FSR, clinical data during routine stroke care in the hospitals were recorded along with baseline information on variables such as demographics, prior history, comorbidity, and functional level

before stroke onset. The definitions of these variables have been previously described [28,29].

Stroke Prognostic Scores

The conventional stroke prognostic scores were used for comparison against data-driven prediction models. In this study, we selected prognostic scores based on the following criteria: they are multiitem and point-based scores using demographic and clinical information, they were developed to predict short-term outcomes after ischemic stroke, and they were externally validated. Consequently, 3 stroke prognostic scores, the ASTRAL score [6], PLAN score [7], and iScore [4,5], were used for comparative analysis. Items of these preexisting stroke prognostic scores were used as explanatory variables in item-based models (Multimedia Appendix 1).

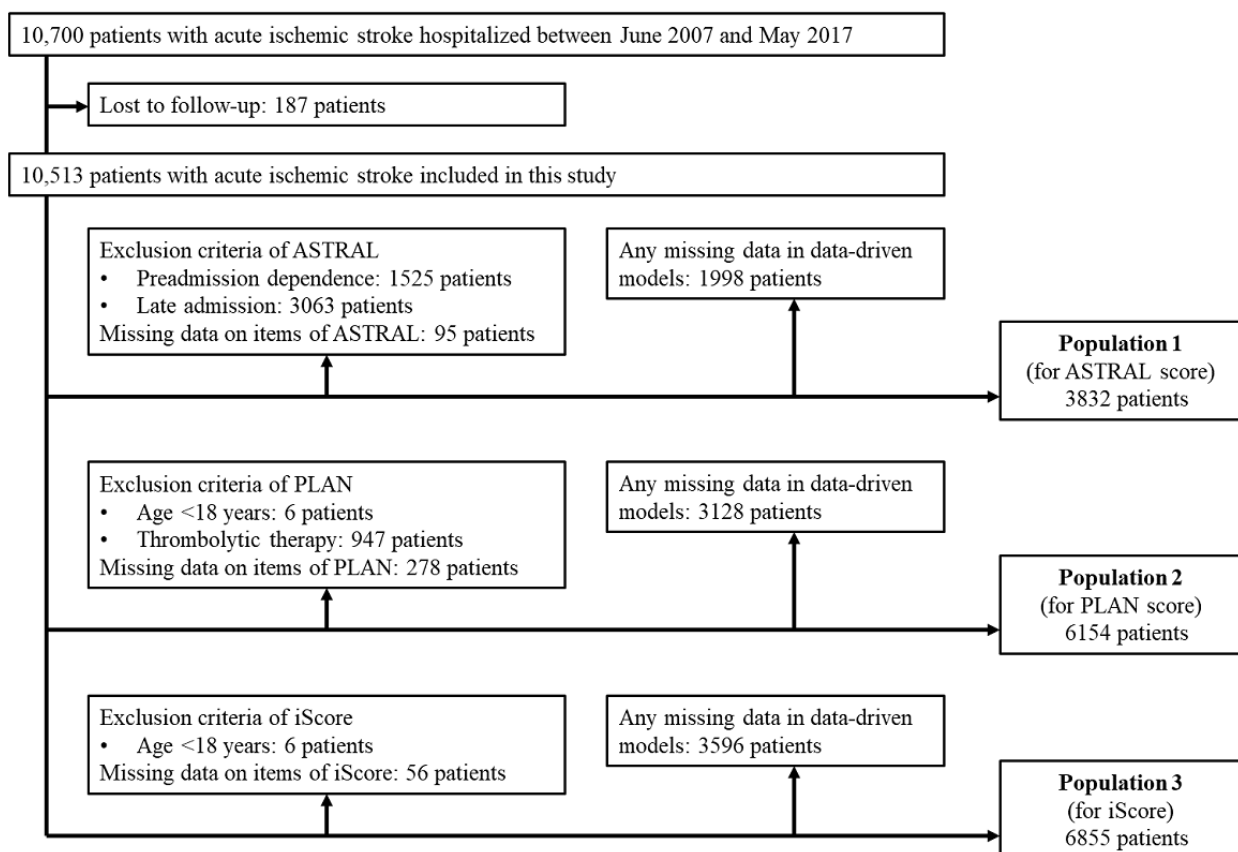
Study Populations

FSR included 10,700 consecutive patients with acute ischemic stroke who were registered between June 2007 and May 2017.

Ischemic stroke was diagnosed based on the sudden onset of a nonconvulsive and focal neurological deficit confirmed by brain imaging through computed tomography, magnetic resonance imaging, or both conducted upon admission. Of the 10,700 patients, 187 (1.7%) were lost to follow-up, and the remaining 10,513 (98.3%) were analyzed for 3 months post stroke.

Study patients were selected according to the inclusion and exclusion criteria of preexisting stroke prognostic scores to make the study populations identical between the item-based and machine learning-based models (Multimedia Appendix 2). Furthermore, we limited the study to patients with complete data, ensuring there were no missing variables across all data points. This approach aimed to prevent further reduction in the number of analyzed patients owing to list-wise deletion in regression models. The frequency of missing data is shown in Multimedia Appendix 3. Consequently, population 1, population 2, and population 3 were included in the analysis for comparison with the ASTRAL score, PLAN score, and iScore, respectively. Figure 1 illustrates the patient selection in each population.

Figure 1. Flowchart for the selection of study patients. Study patients were selected according to the inclusion and exclusion criteria used in the original studies of 3 stroke prognostic scores: population 1 for the Acute Stroke Registry and Analysis of Lausanne (ASTRAL) score, population 2 for the preadmission comorbidities, level of consciousness, age, and neurologic deficit (PLAN) score, and population 3 for the Ischemic Stroke Predictive Risk Score (iScore). Patients with missing data on explanatory variables were excluded from the analyses of data-driven models to avoid the influence of list-wise deletion.



Study Outcomes

The study outcomes were poor functional outcome and death at 3 months after stroke. Poor functional outcome was defined as a modified Rankin Scale score >2 at 3 months after stroke onset [30]. Death was defined as death from any cause within

3 months after stroke [30]. Interviewers on clinical outcomes were blinded to the patients' backgrounds.

Development of Predictive Models

We performed logistic regression analysis to develop item-based models using the predictors of the ASTRAL score, PLAN score,

and iScore as explanatory variables ([Multimedia Appendix 1](#)). The predictors used in these models included age, time delay from onset to admission, stroke scale score, decreased level of consciousness, visual field defect, and abnormal glucose levels for the ASTRAL score; age, atrial fibrillation, congestive heart failure, cancer, preadmission dependence, decreased level of consciousness, leg weakness, arm weakness, and aphasia or neglect for the PLAN score; age, male sex, atrial fibrillation, congestive heart failure, renal dialysis, cancer, preadmission dependence, Canadian Neurological Scale score, stroke subtype, and abnormal glucose levels for the iScore. The categorization of predictors in the stroke prognostic scores was the same as that used in the original study for each score.

We used regularization methods (ridge regression [RR] and least absolute shrinkage and selection operator [LASSO] regression models) and ensemble decision tree models (random forest [RF] and Extreme Gradient Boosting [XGBoost]) for data-driven models based on machine learning algorithms [31-34]. All available variables were included in the development of data-driven models ([Multimedia Appendix 3](#)). The details of the model development are presented in [Multimedia Appendix 4](#).

Metrics of Model Performance

The discriminative ability of each model was evaluated using the area under the receiver operating characteristic curve (AUROC) and the area under the precision-recall curve (AUPRC). AUPRC was calculated because it is a useful performance metric for unbalanced data of infrequent outcome events, such as death [35].

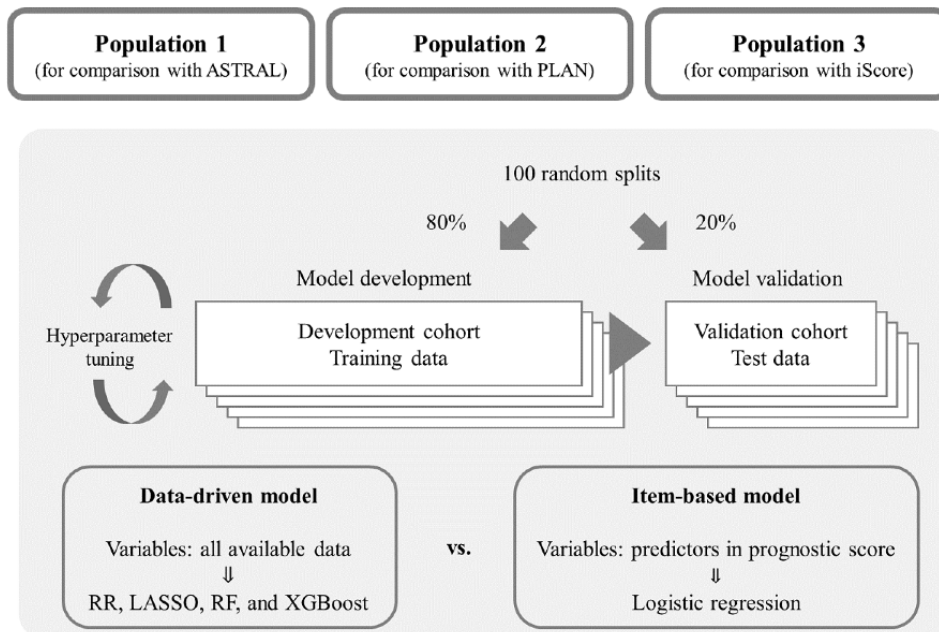
The calibration of each model was assessed using a calibration plot. Calibration plots were obtained by plotting the predicted

and observed probabilities of the clinical outcomes in the 10 risk groups estimated using each predictive model. The Brier score was also used to assess the overall performance. The Brier score is defined as $1/N \sum_{i=1}^N (p_i - a_i)^2$, ($0 \leq BS \leq 1$), where p_i is the predicted probability of the occurrence of an event ranging from 0 to 1, a_i indicates the event with binary outcomes (1 for observed or 0 for not observed), and N is the number of samples.

Validation and Comparison of Models

We performed internal validation of item-based and data-driven models after 100 repeated random splits into 80% of the patients as a training set and 20% of patients as a test set ([Figure 2](#)). The parameters in the training set were optimally tuned via 10-fold cross-validation in the data-driven models. After 100 random splits, the predictive models were developed by logistic regression using the items of the stroke prognostic scores (item-based model) and by machine learning using all variables (data-driven model) in the training set. The developed item-based and data-driven models were validated in the test set. The data sets for both training and testing were identical for the item-based and data-driven models. The median and 95% CI of the performance metrics, that is, AUROC, AUPRC, and Brier score, were calculated for each model using the results of the 100 repeated random splits. To directly compare the performance of the item-based and data-driven models (RR, LASSO, RF, and XGBoost), we compared the AUROC, AUPRC, and Brier score of the data-driven models with those of the corresponding item-based model. We repeated the comparison 100 times and calculated the times that the AUROC, AUPRC, and Brier score of data-driven models were better than those of the corresponding item-based model among the 100 repetitions.

Figure 2. Schematic diagram of the development and validation of the predictive models. All patients were randomly split into 80% of the development cohort as training data and 20% of the validation cohort as test data, which was repeated 100 times. Among the data-driven models, predictive models were developed based on ridge regression (RR), least absolute shrinkage and selection operator regression (LASSO), random forest (RF), and Extreme Gradient Boosting (XGBoost) using all available data after hyperparameter tuning in the development cohort. Logistic regression was used with predictors of stroke prognostic scores in the item-based models. The predictive models were validated using the test data of the validation cohort. In each split, the training and test data were identical between the data-driven and item-based models. ASTRAL: Acute Stroke Registry and Analysis of Lausanne; PLAN: preadmission comorbidities, level of consciousness, age, and neurologic deficit.



Evaluation of the Contribution of Variables

We evaluated the importance of the variables used in the item-based and data-driven models. To assess the contribution of each predictor to the item-based regression model, we calculated the rate of times when the association between each variable and clinical outcomes was statistically significant ($P < .05$) after 100 random splits. In the machine learning models, the magnitude of variable importance was evaluated in identical populations after 100 random splits ([Multimedia Appendix 4](#)).

We calculated the AUROC of the XGBoost model using various types of variables to assess how the addition of explanatory variables improves the predictive performance of the data-driven model. First, we constructed a model with age, sex, National Institutes of Health Stroke Scale (NIHSS) score, and preadmission modified Rankin Scale score (model 1). Then, 5 models were developed by adding items relating to preadmission status to model 1 (model 2), items relating to clinical data on admission to model 2 (model 3), items relating to brain imaging data to model 3 (model 4), and items relating to laboratory data to model 4 (model 5).

Statistical Analysis

We used the chi-square test, 2-tailed Student t test, or Mann-Whitney U test to compare the differences in baseline characteristics and clinical data, as appropriate [36]. Two-sided P values $< .05$ were considered statistically significant.

All statistical analyses were performed using the R statistical package (R Development Core Team). This study was conducted in accordance with the Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD) initiative [37].

Results

Baseline Variables and Clinical Outcomes

The mean age of the 10,513 patients was 73.0 (SD 12.5) years, and 59.1% (6209/10,513) of the patients were men. At 3 months after stroke, a poor functional outcome was found in 1204 (31.4%) of 3832 patients in population 1, 2209 (35.9%) of 6154 patients in population 2, and 2540 (37.1%) of 6855 patients in population 3. Within 3 months after stroke onset, 3% (113/3832), 3.6% (219/6154), and 3.7% (255/6855) of the patients died in population 1, population 2, and population 3, respectively.

First, we investigated the differences in the predictors of preexisting point-based stroke prognostic scores among patients according to poststroke clinical outcomes. Consequently, almost all variables significantly ($P < .05$) differed depending on the 3-month functional outcome ([Table 1](#)) and 3-month survival status ([Multimedia Appendix 5](#)) in addition to the predictors used in preexisting prognostic scores.

Table 1. Baseline data according to functional outcome at 3 months.

	Overall (n=10,513)	mRS ^a 0-2 (n=6405)	mRS 3-6 (n=4108)	P value
Demographics				
Age (y), mean (SD)	73.0 (12.5)	68.9 (12.0)	79.4 (10.4)	<.001
Men, n (%)	6209 (59.1)	4257 (66.5)	1952 (47.5)	<.001
Risk factors, n (%)				
Hypertension	8485 (80.7)	5138 (80.2)	3347 (81.5)	.11
Diabetes mellitus	3607 (34.3)	2236 (34.9)	1371 (33.4)	.11
Atrial fibrillation	2743 (26.1)	1173 (18.3)	1570 (38.3)	<.001
Smoking	2261 (23.1)	1717 (28.9)	544 (14.2)	<.001
Comorbid conditions, n (%)				
Congestive heart failure	919 (8.7)	423 (6.6)	496 (12.1)	<.001
Kidney disease on dialysis	332 (3.2)	171 (2.7)	161 (3.9)	<.001
Cancer	1552 (14.8)	774 (12.1)	778 (18.9)	<.001
Previous history, n (%)				
Previous myocardial infarction	505 (5.3)	242 (4.3)	263 (6.9)	<.001
Preadmission functional status				
Preadmission mRS, median (IQR)	0 (0-1)	0 (0-0)	1 (0-3)	<.001
Preadmission dependence (mRS score >1), n (%)	2366 (22.5)	364 (5.7)	2002 (48.7)	<.001
Onset-to-admission time, n (%)				
≤1 h	943 (9)	490 (7.7)	453 (11)	<.001
≤3 h	1469 (14)	771 (12)	698 (17)	<.001
≤6 h	1141 (10.9)	644 (10.1)	497 (12.1)	<.001
≤24 h	3515 (33.4)	2090 (32.6)	1425 (34.7)	<.001
>24 h	3445 (32.8)	2410 (37.6)	1035 (25.2)	<.001
Stroke subtype, n (%)				
Small vessel occlusion	2119 (20.2)	1724 (26.9)	395 (9.6)	<.001
Large artery atherosclerosis	1823 (17.3)	1006 (15.7)	817 (19.9)	<.001
Cardioembolism	2496 (23.7)	1054 (16.5)	1442 (35.1)	<.001
Other determined etiology	2146 (20.4)	1404 (21.9)	742 (18.1)	<.001
Undetermined	1929 (18.3)	1217 (19)	712 (17.3)	<.001
Neurological severity, median (IQR) or n (%)				
NIHSS ^b score	3 (2-8)	2 (1-4)	8 (4-16)	<.001
Severe stroke (NIHSS score >10)	1938 (18.4)	291 (4.5)	1647 (40.1)	<.001
Neurological deficits, n (%)				
Decreased level of consciousness	3129 (30)	770 (12.1)	2359 (57.9)	<.001
Leg weakness	5394 (51.9)	2357 (37.2)	3037 (75)	<.001
Arm weakness	5634 (54.2)	2520 (39.7)	3114 (76.8)	<.001
Aphasia or neglect	2912 (27.9)	946 (14.9)	1966 (48.3)	<.001
Visual field defect	999 (9.6)	447 (7.0)	552 (13.6)	<.001
Physiological data, mean (SD)				
SBP ^c , mm Hg	86.6 (18.2)	87.9 (17.8)	84.6 (18.6)	<.001
DBP ^d , mm Hg	159.8 (29.3)	160.4 (28.6)	158.8 (30.3)	.01

	Overall (n=10,513)	mRS ^a 0-2 (n=6405)	mRS 3-6 (n=4108)	P value
BMI, kg/m ²	22.8 (3.8)	23.5 (3.6)	21.7 (3.9)	<.001
Laboratory data, median (IQR)				
Complete blood cell count				
WBC ^e (10 ³ /μL)	6.8 (5.6-8.4)	6.7 (5.5-8.2)	7.0 (5.7-8.9)	<.001
RBC ^f (10 ⁴ /μL)	436 (394-476)	449 (411-485)	416 (372-458)	<.001
Hematocrit (%)	40.1 (36.5-43.4)	41.1 (37.9-44.0)	38.2 (34.6-41.9)	<.001
Hemoglobin (g/dL)	13.5 (12.1-14.8)	14.0 (12.7-15.1)	12.8 (11.4-14.1)	<.001
Platelet (10 ⁴ /μL)	20.2 (16.6-24.3)	20.6 (17.0-24.7)	19.5 (15.8-23.6)	<.001
Liver function				
AST ^g (U/L)	23 (19-29)	23 (19-29)	23 (19-30)	.001
ALT ^h (U/L)	17 (12-24)	18 (13-25)	15 (11-22)	<.001
LDH ⁱ (U/L)	219 (186-266)	211 (181-254)	230 (195-285)	<.001
ALP ^j (U/L)	239 (195-295)	231 (190-284)	250 (203-312)	<.001
Kidney function				
BUN ^k (mg/dL)	16.0 (13.0-20.9)	15.3 (12.6-19.0)	17.9 (13.8-23.8)	<.001
Creatinine (mg/dL)	0.8 (0.6-1.0)	0.8 (0.7-1.0)	0.8 (0.6-1.1)	<.001
eGFR ^l (mL/min/1.73 m ²)	66.5 (51.2-81.5)	70.2 (55.9-83.8)	60.8 (44.8-76.5)	<.001
Glycemic control				
Glucose (mg/100 mL)	121 (103-156)	119 (103-154)	124 (105-158)	.001
Hemoglobin A _{1c} (%)	5.9 (5.6-6.6)	5.9 (5.6-6.6)	5.9 (5.5-6.5)	<.001
Inflammation				
hsCRP ^m , mg/dL	1.5 (0.5-6.1)	1.0 (0.4-2.9)	3.9 (1.0-16.3)	<.001
Coagulation				
PT-INR ⁿ	1.0 (1.0-1.1)	1.0 (1.0-1.1)	1.1 (1.0-1.1)	<.001
APTT ^o (s)	29.7 (27.2-32.7)	29.5 (27.1-32.4)	30.1 (27.3-33.3)	<.001
Fibrinogen (mg/dL)	304 (260-359)	297 (256-349)	315 (267-375)	<.001
d-dimer (μg/mL)	0.9 (0.4-2.0)	0.6 (0.2-1.2)	1.7 (0.9-4.0)	<.001

^amRS: modified Rankin Scale.

^bNIHSS: National Institutes of Health Stroke Scale.

^cSBP: systolic blood pressure.

^dDBP: diastolic blood pressure.

^eWBC: white blood cell count.

^fRBC: red blood cell count.

^gAST: aspartate aminotransferase.

^hALT: alanine aminotransferase.

ⁱLDH: lactate dehydrogenase.

^jALP: alkaline phosphatase.

^kBUN: blood urea nitrogen.

^leGFR: estimated glomerular filtration rate.

^mhsCRP: high-sensitivity C-reactive protein.

ⁿPT-INR: international normalized ratio of prothrombin time.

^oAPTT: activated partial thromboplastin time.

Assessment of Model Performance

AUROC varied depending on study populations, whereas differences between the machine learning algorithms were minimal in the same study population and for the same outcome. The AUROCs of data-driven models based on machine learning were generally higher than those of item-based models for

predicting both 3-month poor functional outcome and all-cause death (Table 2). Similarly, AUPRCs were generally higher in data-driven models than in item-based models for predicting both poor functional outcome and all-cause death (Table 3). Regarding the Brier score, the data-driven models performed better than the item-based models (Table 4).

Table 2. Area under the receiver operating characteristic curve for predicting unfavorable clinical outcomes at 3 months using item-based and data-driven models^a.

	Item-based model, median (95% CI)	Data-driven models, median (95% CI)			
		RR ^b	LASSO ^c	RF ^d	XGBoost ^e
Poor functional outcome					
Population 1 (n=3832)	0.83 (0.80-0.85)	0.86 (0.83-0.89)	0.86 (0.84-0.89)	0.86 (0.84-0.88)	0.86 (0.83-0.89)
Population 2 (n=6154)	0.88 (0.86-0.90)	0.91 (0.90-0.93)	0.91 (0.90-0.93)	0.91 (0.89-0.92)	0.91 (0.89-0.93)
Population 3 (n=6855)	0.87 (0.85-0.89)	0.90 (0.89-0.92)	0.90 (0.89-0.92)	0.90 (0.88-0.91)	0.90 (0.89-0.92)
Death					
Population 1 (n=3832)	0.77 (0.69-0.87)	0.87 (0.79-0.93)	0.87 (0.78-0.92)	0.89 (0.81-0.93)	0.88 (0.82-0.93)
Population 2 (n=6154)	0.84 (0.80-0.89)	0.89 (0.85-0.92)	0.88 (0.84-0.92)	0.90 (0.86-0.93)	0.90 (0.86-0.93)
Population 3 (n=6855)	0.82 (0.77-0.87)	0.88 (0.84-0.91)	0.87 (0.83-0.90)	0.89 (0.86-0.92)	0.89 (0.85-0.91)

^aThe study populations were selected according to the inclusion and exclusion criteria for the Acute Stroke Registry and Analysis of Lausanne (ASTRAL) score (population 1), the preadmission comorbidities, level of consciousness, age, and neurologic deficit (PLAN) score (population 2), and the Ischemic Stroke Predictive Risk Score (iScore; population 3).

^bRR: ridge regression.

^cLASSO: least absolute shrinkage and selection operator regression.

^dRF: random forest.

^eXGBoost: Extreme Gradient Boosting.

Table 3. Area under the precision-recall curve for predicting unfavorable clinical outcomes at 3 months using item-based and data-driven models^a.

	Item-based model, median (95% CI)	Data-driven models, median (95% CI)			
		RR ^b	LASSO ^c	RF ^d	XGBoost ^e
Poor functional outcome					
Population 1 (n=3832)	0.71 (0.66-0.75)	0.75 (0.71-0.79)	0.75 (0.71-0.80)	0.74 (0.69-0.79)	0.75 (0.71-0.79)
Population 2 (n=6154)	0.83 (0.80-0.86)	0.87 (0.85-0.89)	0.87 (0.85-0.90)	0.87 (0.84-0.89)	0.87 (0.85-0.89)
Population 3 (n=6855)	0.83 (0.80-0.85)	0.87 (0.85-0.89)	0.87 (0.85-0.89)	0.86 (0.84-0.88)	0.87 (0.85-0.89)
Death					
Population 1 (n=3832)	0.11 (0.06-0.24)	0.17 (0.08-0.32)	0.17 (0.07-0.31)	0.26 (0.13-0.44)	0.24 (0.12-0.39)
Population 2 (n=6154)	0.17 (0.11-0.25)	0.27 (0.18-0.37)	0.27 (0.18-0.38)	0.29 (0.18-0.42)	0.27 (0.16-0.35)
Population 3 (n=6855)	0.18 (0.11-0.25)	0.27 (0.16-0.36)	0.27 (0.17-0.38)	0.29 (0.19-0.42)	0.28 (0.19-0.39)

^aThe study populations were selected according to the inclusion and exclusion criteria for the Acute Stroke Registry and Analysis of Lausanne (ASTRAL) score (population 1), the preadmission comorbidities, level of consciousness, age, and Neurologic deficit (PLAN) score (population 2), and the Ischemic Stroke Predictive Risk Score (iScore; population 3).

^bRR: ridge regression.

^cLASSO: least absolute shrinkage and selection operator regression.

^dRF: random forest.

^eXGBoost: Extreme Gradient Boosting.

Table 4. Brier score for predicting unfavorable clinical outcomes at 3 months using item-based and data-driven models^a.

	Item-based model, median (95% CI)	Data-driven models, median (95% CI)			
		RR ^b	LASSO ^c	RF ^d	XGBoost ^e
Poor functional outcome					
Population 1 (n=3832)	0.15 (0.14-0.17)	0.14 (0.12-0.15)	0.14 (0.12-0.15)	0.14 (0.13-0.15)	0.14 (0.12-0.15)
Population 2 (n=6154)	0.13 (0.12-0.14)	0.11 (0.10-0.12)	0.11 (0.10-0.12)	0.12 (0.11-0.13)	0.11 (0.10-0.12)
Population 3 (n=6855)	0.13 (0.12-0.15)	0.12 (0.11-0.13)	0.12 (0.11-0.13)	0.12 (0.12-0.13)	0.12 (0.11-0.13)
Death					
Population 1 (n=3832)	0.03 (0.02-0.03)	0.03 (0.02-0.03)	0.03 (0.02-0.03)	0.03 (0.02-0.03)	0.03 (0.02-0.03)
Population 2 (n=6154)	0.03 (0.02-0.04)	0.03 (0.02-0.04)	0.03 (0.02-0.04)	0.03 (0.02-0.04)	0.03 (0.02-0.04)
Population 3 (n=6855)	0.03 (0.02-0.04)	0.03 (0.02-0.04)	0.03 (0.02-0.04)	0.03 (0.02-0.04)	0.03 (0.02-0.04)

^aThe study populations were selected according to the inclusion and exclusion criteria for the Acute Stroke Registry and Analysis of Lausanne (ASTRAL) score (population 1), the preadmission comorbidities, level of consciousness, age, and neurologic deficit (PLAN) score (population 2), and the Ischemic Stroke Predictive Risk Score (iScore; population 3).

^bRR: ridge regression.

^cLASSO: least absolute shrinkage and selection operator regression.

^dRF: random forest.

^eXGBoost: Extreme Gradient Boosting.

The predictive performance of data-driven models compared with the corresponding item-based model was examined by the frequency of the performance metrics (AUROC, AUPRC, and Brier score) of data-driven models, which were better than those of the corresponding item-based model in the identical training and test data sets after 100 repeated random splits (Table 5). Regarding poor functional outcome, the frequency exceeded 95% for all metrics in all the data-driven models (RR, LASSO, RF, and XGBoost), indicating that the probability of the worse performance of data-driven models compared with the item-based model was <5%. Regarding death, the frequency

was >95% for AUROC in all the data-driven models but did not always attain 95% for AUPRC or Brier score.

Calibration for predicting poor functional outcome was compared between the item-based and data-driven models (RR, LASSO, RF, and XGBoost) in population 1 for the ASTRAL score, in population 2 for the PLAN score, and in population 3 for the iScore. The prediction of poor functional outcome (Figure 3) and all-cause death (Figure 4) demonstrated concordance between the predicted and observed probabilities in the item-based models as well as in the data-driven models.

Table 5. Predictive performance of data-driven models versus item-based models^a.

	Poor functional outcome				Death			
	RR ^b	LASSO ^c	RF ^d	XGBoost ^e	RR	LASSO	RF	XGBoost
AUROC^f								
Population 1 (n=3832)	100	100	100	100	97	95	97	96
Population 2 (n=6154)	100	100	100	100	100	100	98	99
Population 3 (n=6855)	100	100	100	100	100	99	100	99
AUPRC^g								
Population 1 (n=3832)	100	100	99	98	81	78	93	93
Population 2 (n=6154)	100	100	100	100	99	99	99	100
Population 3 (n=6855)	100	100	100	100	98	98	100	98
Brier score								
Population 1 (n=3832)	100	100	99	100	83	70	96	89
Population 2 (n=6154)	100	100	100	100	98	92	97	93
Population 3 (n=6855)	100	100	100	100	100	99	100	96

^aData indicate the frequency that AUROC, AUPRC, and Brier score of data-driven models (RR, LASSO, RF, or XGBoost) exceeded those of item-based models in identical training and test sets after 100 repeated random splits.

^bRR: ridge regression.

^cLASSO: least absolute shrinkage and selection operator regression.

^dRF: random forest.

^eXGBoost: Extreme Gradient Boosting.

^fAUROC: area under the receiver operating characteristic curve.

^gAUPRC: area under the precision-recall curve.

Figure 3. Calibration of item-based and data-driven models for predicting poor functional outcome. Calibration for predicting poor functional outcome was compared between the item-based regression model and data-driven models (ridge regression [RR], least absolute shrinkage and selection operator regression [LASSO], random forest [RF], and Extreme Gradient Boosting [XGBoost]) in population 1 for the Acute Stroke Registry and Analysis of Lausanne (ASTRAL) score, population 2 for the preadmission comorbidities, level of consciousness, age, and neurologic deficit (PLAN) score, and population 3 for the Ischemic Stroke Predictive Risk Score (iScore). The patients were categorized into 10 groups stratified by the predicted probability of poor functional outcome in the test data. Observed probabilities (x-axis) were plotted against predicted probabilities (y-axis) in the 10 groups based on risk stratification. The results for the first 100 random splits are presented.

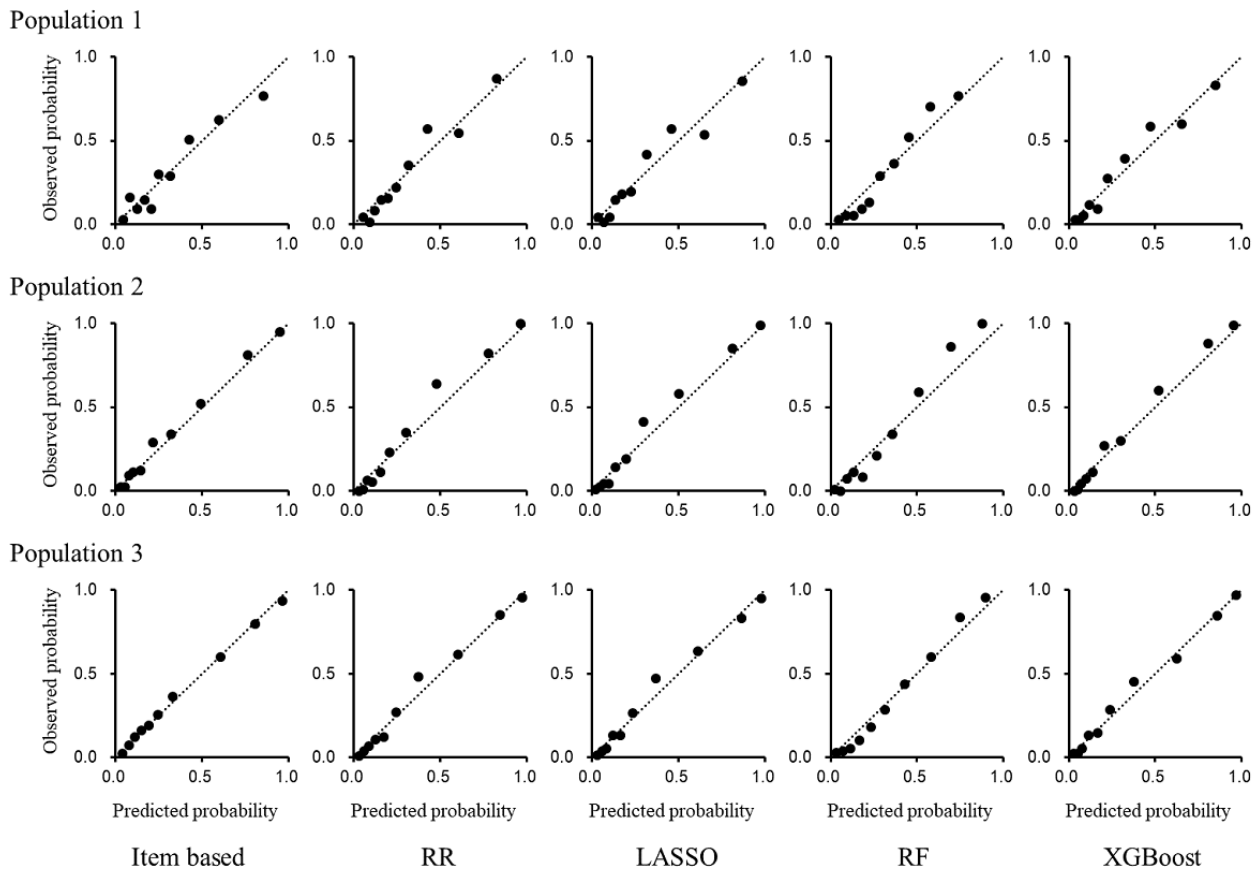
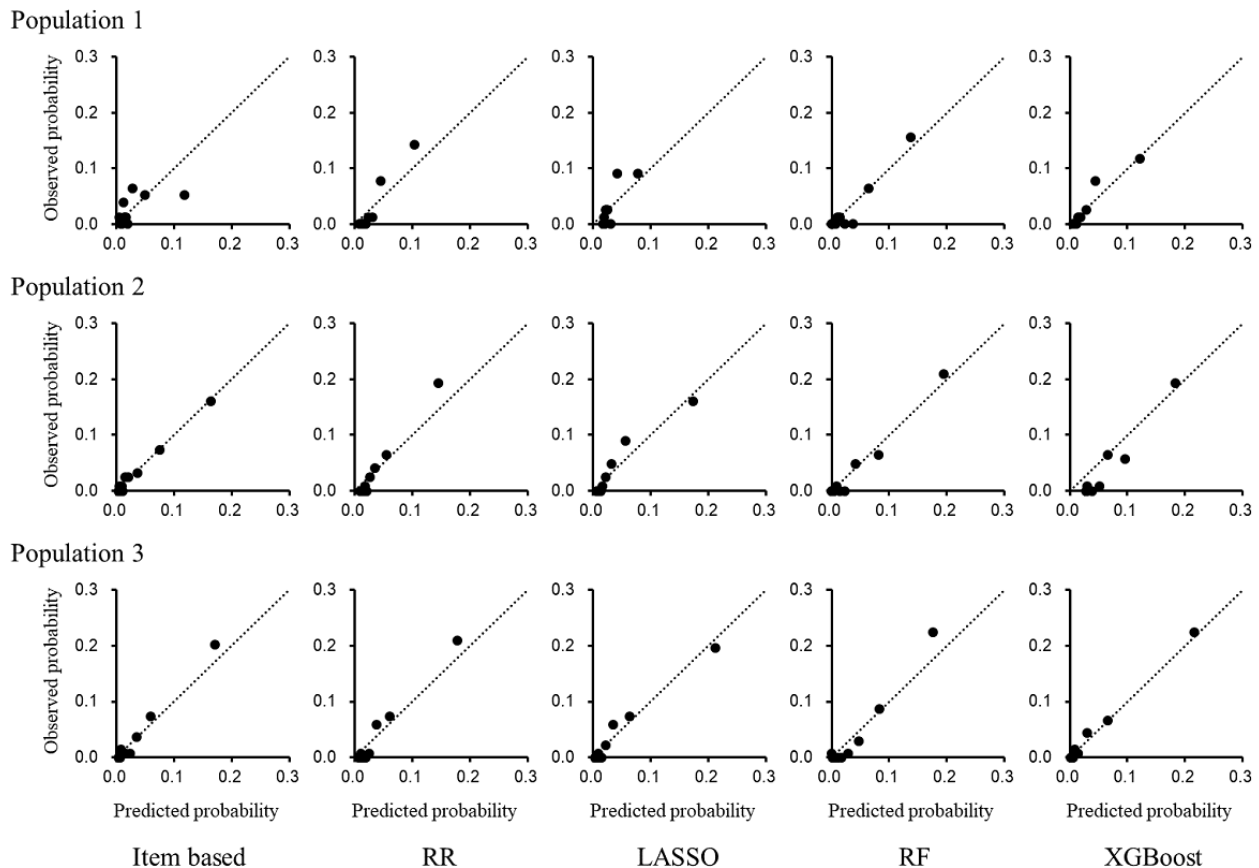


Figure 4. Calibration of item-based and data-driven models for predicting death. Calibration for predicting death was compared between the item-based regression model and data-driven models (ridge regression [RR], least absolute shrinkage and selection operator regression [LASSO], random forest [RF], and Extreme Gradient Boosting [XGBoost]) in population 1 for the Acute Stroke Registry and Analysis of Lausanne (ASTRAL) score, population 2 for the preadmission comorbidities, level of consciousness, age, and neurologic deficit (PLAN) score, and population 3 for the Ischemic Stroke Predictive Risk Score (iScore). The patients were categorized into 10 groups stratified by the predicted probability of death in the test data. Observed probabilities (x-axis) were plotted against predicted probabilities (y-axis) in the 10 groups based on risk stratification. The results for the first 100 random splits are presented.



Evaluation of Variables

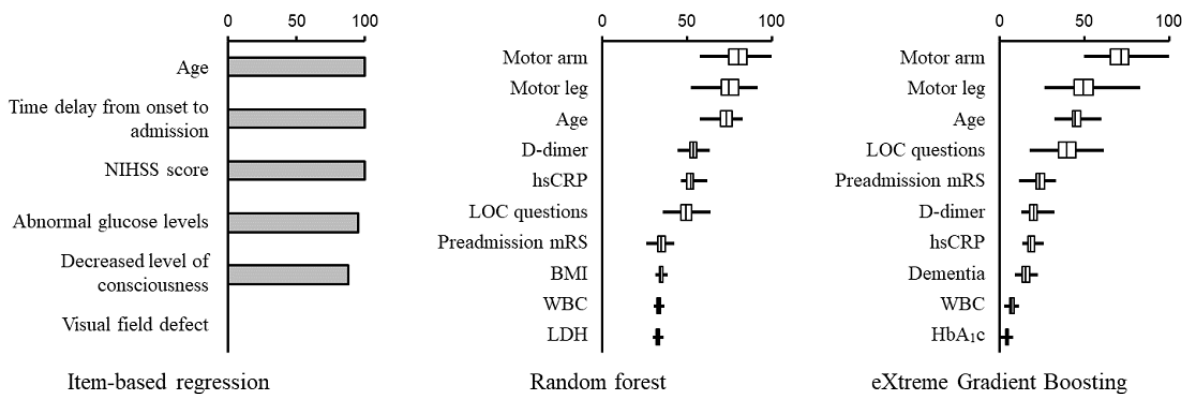
Next, we evaluated how each variable contributed to the predictive performance of the item-based and data-driven models (RF and XGBoost) in population 1 (Figure 5), population 2 (Figure 6), and population 3 (Figure 7). The selected variables differed substantially between the study populations in the item-based models. Age, preadmission dependence, and neurological severity of stroke were important variables in predicting both poor functional outcome and death (Figures 5-7; left panels). Age and neurological deficit signs (arm or leg weakness and loss of consciousness) were the most frequently used variables for predicting poor functional outcome (Figures 5A, 6A, and 7A; middle and right panels) in RF and XGBoost.

In contrast, variables not used in the item-based models, such as d-dimer, high-sensitivity C-reactive protein, fibrinogen, and BMI, were the most frequently used variables by RF and XGBoost (Figures 5B, 6B, and 7B; middle and right panels) in predicting death.

We also investigated how the addition of variables increased the predictive performance of XGBoost. As a result, the AUROC for poor functional outcome did not substantially increase even when explanatory variables other than key predictors were added to model 1 (Figure 8; open circles). Conversely, the AUROC for all-cause death linearly increased with the addition of other variables to the models, particularly items from laboratory data (Figure 8; closed circles).

Figure 5. Comparison of variable importance between items of the Acute Stroke Registry and Analysis of Lausanne (ASTRAL) score and explanatory variables in machine learning model in population 1. The contribution of each variable to the models in predicting poor functional outcome (A) and death (B) is shown. The patients were selected based on the ASTRAL criteria (population 1). In item-based regression models, the percentage indicates the rate of times when its association with clinical outcomes was statistically significant ($P < .05$). In machine learning models, the top 10 variables are shown according to the magnitude of variable importance. Boxes, vertical lines in the boxes, and horizontal bars indicate IQR, median, and minimal or maximal range, respectively. NIHSS: National Institutes of Health Stroke Scale, hsCRP: high-sensitivity C-reactive protein, LOC: loss of consciousness, mRS: modified Rankin Scale, BMI: body mass index, WBC: white blood cell count, LDH: lactate dehydrogenase, HbA1c: hemoglobin A1c, Fib: fibrinogen, Plt: platelet count, RBC: red blood cell count, ALP: alkaline phosphatase, Ht: hematocrit, Hb: hemoglobin, BUN: blood urea nitrogen, LDH: lactate dehydrogenase, PT-INR: international normalized ratio of prothrombin time.

A



B

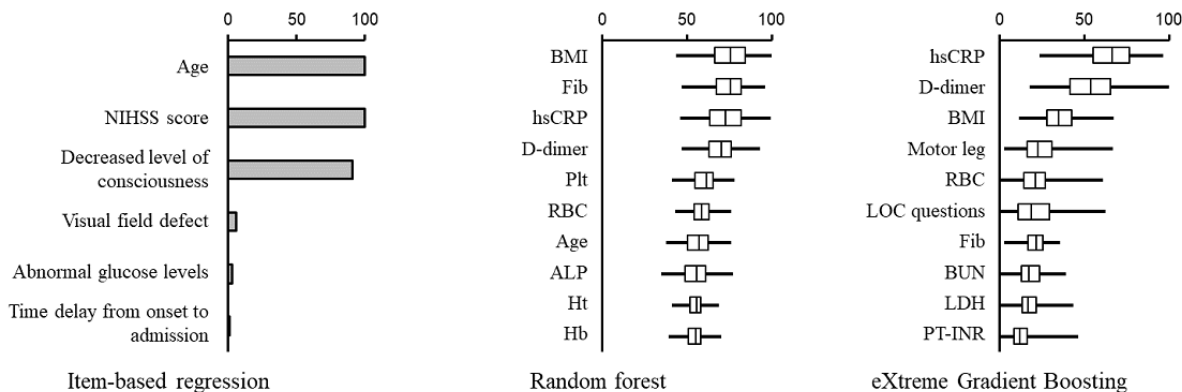
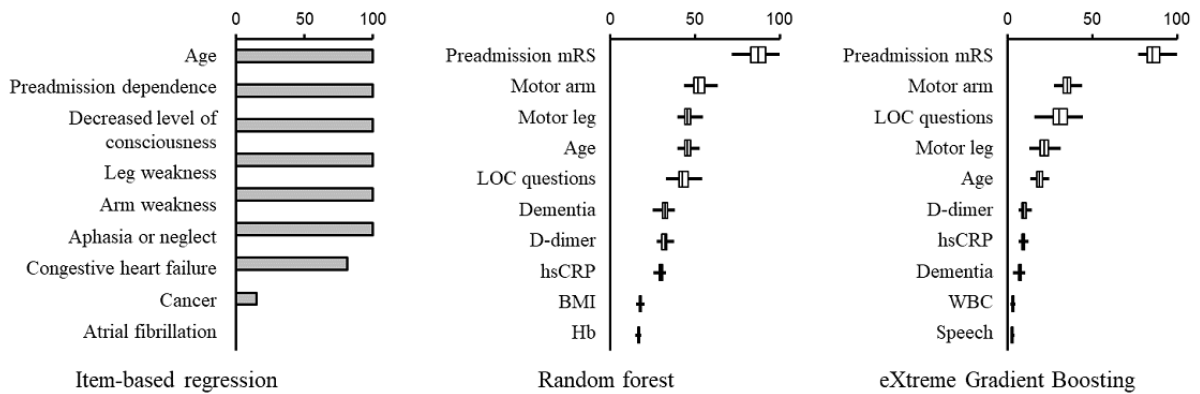


Figure 6. Comparison of variable importance between items of the preadmission comorbidities, level of consciousness, age, and neurologic deficit (PLAN) score and explanatory variables in machine learning model in population 2. The contribution of each variable to the models in predicting poor functional outcome (A) and death (B) is shown. The patients were selected based on the PLAN score criteria (population 2). In item-based regression models, the percentage indicates the rate of times when its association with clinical outcomes was statistically significant ($P < .05$). In machine learning models, the top 10 variables are shown according to the magnitude of variable importance. Boxes, vertical lines in the boxes, and horizontal bars indicate IQR, median, and minimal or maximal range, respectively. mRS: modified Rankin Scale, LOC: loss of consciousness, hsCRP: high-sensitivity C-reactive protein, BMI: body mass index, Hb: hemoglobin, WBC: white blood cell count, Plt: platelet count, Fib: fibrinogen, RBC: red blood cell count, LDH: lactate dehydrogenase, Ht: hematocrit, ALP: alkaline phosphatase, PT-INR: international normalized ratio of prothrombin time.

A



B

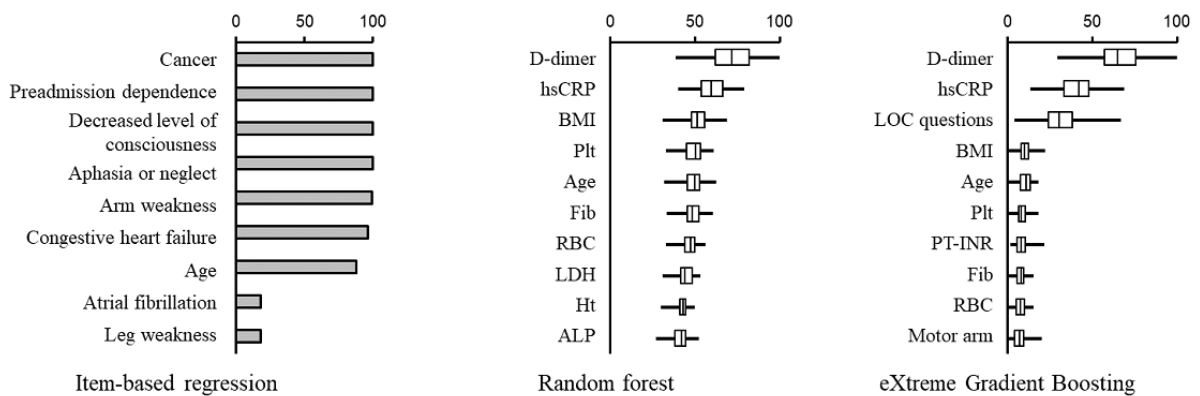


Figure 7. Comparison of variable importance between items of Ischemic Stroke Predictive Risk Score (iScore) and explanatory variables in machine learning model in population 3. The contribution of each variable to the models in predicting poor functional outcome (A) and death (B) is shown. The patients were selected according to the iScore criteria (population 3). In item-based regression models, the percentage indicates the rate of times when its association with clinical outcomes was statistically significant ($P < .05$). In machine learning models, the top 10 variables are shown according to the magnitude of variable importance. Boxes, vertical lines in the boxes, and horizontal bars indicate IQR, median, and minimal or maximal range, respectively. NIHSS: National Institutes of Health Stroke Scale, CNS: Canadian Neurological Scale, mRS: modified Rankin Scale, LOC: loss of consciousness, hsCRP: high-sensitivity C-reactive protein, BMI: body mass index, Hb: hemoglobin, WBC: white blood cell count, Fib: fibrinogen, RBC: red blood cell count, Plt: platelet count, Ht: hematocrit, LDH: lactate dehydrogenase, ALP: alkaline phosphatase, PT-INR: international normalized ratio of prothrombin time.

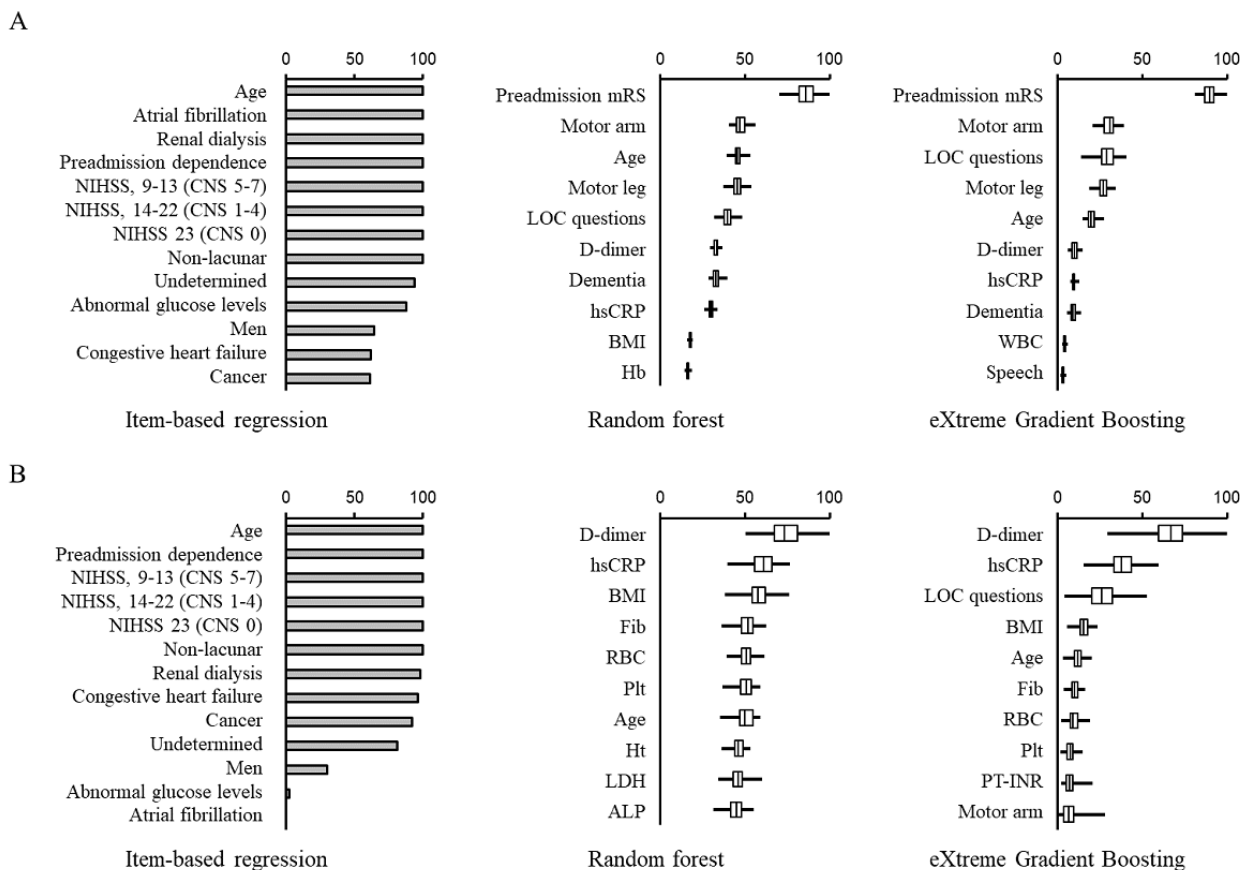
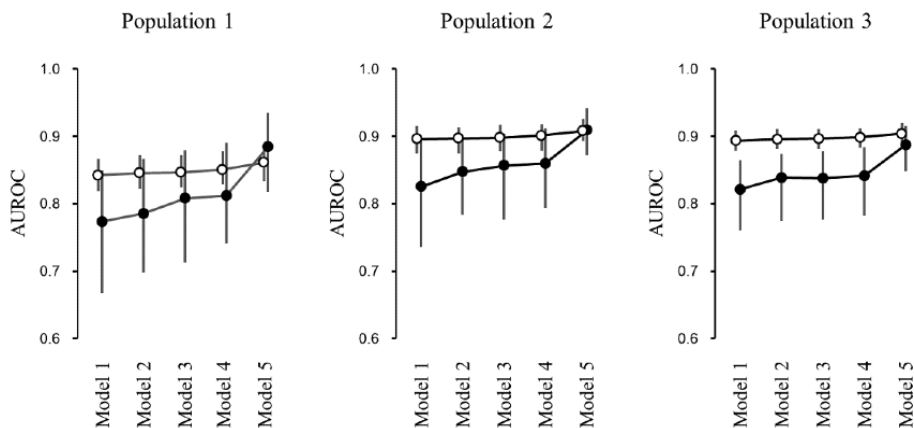


Figure 8. Improvement of discrimination in a data-driven model by adding different types of data. The area under the receiver operating characteristic curves (AUROCs) for predicting poor functional outcome (open circles) and death (closed circles) were compared among the 5 models, which used different types of variables. A data-driven model was developed for each population using Extreme Gradient Boosting. Vertical bars indicate the 95th percentile after 100 random splits. The variables used for the models were as follows: model 1: age, sex, National Institutes of Health Stroke Scale score, and preadmission modified Rankin Scale score; model 2: model 1 plus clinical data before admission (eg, risk factors, comorbid conditions, previous history, family history, and prestroke medication); model 3: model 2 plus clinical data on admission (eg, onset-to-admission time, ambulance use, BMI, and physiological data); model 4: model 3 plus brain imaging data (eg, site of lesion, side of lesion, and stroke subtype); and model 5: model 4 plus laboratory data.



Discussion

Principal Findings

This study, which analyzed comprehensive clinical data from a multicenter, hospital-based stroke registry, yielded the following major findings. The performance of item-based regression models using the predictors of 3 conventional stroke prognostic scores was fair in predicting clinical outcomes at 3 months after ischemic stroke in our cohort, despite differences in clinical and social backgrounds from the original cohorts of scores. Data-driven models based on machine learning algorithms exhibited better performance when compared with item-based models in identical study populations. The importance of variables in RF and XGBoost appeared to differ from that in item-based models when predicting death within 3 months. The addition of nonconventional factors, such as laboratory data, to the XGBoost model improved its predictive ability for 3-month mortality.

Predictive Performance of Models

Thus far, only a limited number of studies have evaluated the predictive performance of machine learning–based models compared with those of stroke prognostic scores [19,20,23]. All these studies were performed in single-center registries or under specific conditions, such as large vessel occlusion in ischemic stroke. Furthermore, previous studies mainly focused on AUROC for assessing predictive performance, although other metrics, such as measures of calibration, are necessary to fully evaluate the performance of models [38]. This study was conducted using a multicenter registry database and several performance metrics. Our study demonstrated that data-driven models developed using machine learning algorithms can perform reasonably well in predicting the 3-month clinical outcomes of patients with acute ischemic stroke. Generally, data-driven models performed better than conventional prognostic scores when both were compared in identical study populations.

This study also demonstrates that the model performance largely depends on the study populations. The study populations varied in terms of both size and patient characteristics, such as prestroke dependency, time from onset to admission, and use of thrombolytic therapy. The variability in AUROC, AUPRC, and Brier scores between the study populations was as large as that between the models. Moreover, the model performance varied depending on the outcomes to be predicted: AUPRCs were substantially decreased for the prediction of death, which is a less frequent event than the poor functional outcome. These findings underscore the reiterated importance of sample size, the number of outcome events, and data quality of the study cohorts where models are to be developed and validated [25,39,40].

Variables in Models

In this study, age, preadmission dependence, and variables related to neurological deficits were identified as important predictors for the prediction of poor functional outcome in both item-based regression models and data-driven models using RF and XGBoost. These are well-known risk factors for poor

functional outcome and are also used for predicting death in stroke prognostic scores [4,5,7]. However, BMI and items related to laboratory data, such as D-dimer, high-sensitivity C-reactive protein, and fibrinogen, were found to be the most important variables for predicting death in RF and XGBoost. Indeed, the association between poststroke clinical outcomes and markers of inflammation and hypercoagulation has become a recent research topic [41,42]. Machine learning algorithms can be a promising tool to identify novel factors to be considered in making prognoses for stroke because they can maximize the use of data without arbitrary assumptions and procedures.

Clinical Implications

The ability of machine learning to derive a model that best fits the data on a given cohort is appealing for making prognoses. Prognostic scores with prespecified items may not fit all cohorts because heterogeneity must exist between study cohorts in race or ethnic groups, general health conditions, socioeconomic status, and health care systems. In addition, stroke prognostic scores are at risk of getting outdated over time, as advances in stroke care continuously improve clinical outcomes in patients with stroke [43,44]. However, our analysis suggests that the 3 conventional prognostic scores can perform sufficiently well in our cohort, despite the fact that the original studies that developed the scores had patients with different medical backgrounds and during different study periods. This finding demonstrates the robustness of outcome prediction using regression models in terms of generalizability. Furthermore, considering nonlinear and interaction effects might not be crucial for outcome prediction after ischemic stroke, as the simple regression models worked well in our study.

Point-based stroke prognostic scores are convenient and helpful for making prompt decisions at the bedside. Generally, prognostic scores comprise only a handful of variables on which information can be obtained easily. This advantage in the practicability of the prognostic scores is important in acute stroke care settings. Machine learning algorithms require more data than conventional prognostic scores to reach acceptable performance levels [39], and the data required by machine learning algorithms to realize better performance, such as laboratory data, may not always be available, although they can improve the predictive performance of models. Therefore, further studies are needed to fully assess the incremental value of machine learning–based models in daily clinical practice.

Strengths and Limitations

This study has several strengths. We assessed and compared the predictive accuracy of prognostic scores against data-driven models, using information from a multicenter, prospective registry of individuals diagnosed with acute stroke. We were able to use several variables, including laboratory data–related items, owing to the detailed clinical data available in the registry. Moreover, comparisons of models were made using various performance metrics. However, this study has also several limitations. First, the selection of patients may have led to bias, although the inclusion and exclusion criteria were identical to those reported in the original studies of the prognostic scores. Second, there were missing data for the baseline variables and clinical outcomes, which may have also led to selection bias.

Third, the possibility of overfitting cannot be completely ruled out, despite the predictive models constituted by the training set being fitted to the test set. Finally, this study included only patients with acute ischemic stroke who were hospitalized in tertiary care centers in a restricted region of Japan. Generalizability should be assessed in other settings and for other diseases.

Conclusions

This study suggests that data-driven models based on machine learning algorithms can improve predictive performance by using diverse types of variables, such as laboratory data-related items. The clinical outcomes of individual patients can be automatically estimated using machine learning algorithms if

a large amount of data can be directly drawn from electronic health records. This possibility of making automated and personalized prognoses is an appealing property of data-driven prediction. However, the arrangement of an appropriate electronic infrastructure is indispensable for enabling data collection, and the development of such infrastructure requires time and cost. It is worth noting that conventional prognostic scores can achieve sufficient performance in making stroke prognoses with only a limited number of variables. In the near future, it seems feasible to explore the improvement of preexisting prognostic scores by incorporating novel predictors identified by machine learning algorithms, given the significant investment necessary to fully use machine learning.

Acknowledgments

This study was supported by the Japan Society for Promotion of Science KAKENHI (grants JP21H03165, JP21K19648, 21K10330, and JP22K10386) and the Ministry of Health, Labour and Welfare AC Program (grant JPMH21446713). The authors thank all the Fukuoka Stroke Registry investigators and their hospitals for participating in this study and all the clinical research coordinators from the Hisayama Research Institute for Lifestyle Diseases for their help in obtaining informed consent and collecting clinical data. Participating hospitals in the Fukuoka Stroke Registry included Kyushu University Hospital (Fukuoka, Japan), National Hospital Organization Kyushu Medical Center (Fukuoka, Japan), National Hospital Organization Fukuoka-Higashi Medical Center (Koga, Japan), Fukuoka Red Cross Hospital (Fukuoka, Japan), St Mary's Hospital (Kurume, Japan), Steel Memorial Yawata Hospital (Kitakyushu, Japan), and Japan Labor Health and Welfare Organization Kyushu Rosai Hospital (Kitakyushu, Japan). Steering committee and research working group members of the Fukuoka Stroke Registry were Takao Ishitsuka, MD, PhD (Fukuoka Mirai Hospital, Fukuoka, Japan); Setsuro Ibayashi, MD, PhD (Chair, Seiai Rehabilitation Hospital, Onojo, Japan); Kenji Kusuda, MD, PhD (Seiai Rehabilitation Hospital, Onojo, Japan); Kenichiro Fujii, MD, PhD (Japan Seafarers Relief Association Moji Ekisaikai Hospital, Kitakyushu, Japan); Tetsuhiko Nagao, MD, PhD (Safety Monitoring Committee, Seiai Rehabilitation Hospital, Onojo, Japan); Yasushi Okada, MD, PhD (Vice-Chair, National Hospital Organization Kyushu Medical Center, Fukuoka, Japan); Masahiro Yasaka, MD, PhD (National Hospital Organization Kyushu Medical Center, Fukuoka, Japan); Hiroaki Ooboshi, MD, PhD (Fukuoka Dental College Medical and Dental Hospital, Fukuoka, Japan); Takanari Kitazono, MD, PhD (Principal Investigator, Kyushu University, Fukuoka, Japan); Katsumi Irie, MD, PhD (Hakujuji Hospital, Fukuoka, Japan); Tsuyoshi Omae, MD, PhD (Imazu Red Cross Hospital, Fukuoka, Japan); Kazunori Toyoda, MD, PhD (National Cerebral and Cardiovascular Center, Suita, Japan); Hiroshi Nakane, MD, PhD (National Hospital Organization Fukuoka-Higashi Medical Center, Koga, Japan); Masahiro Kamouchi, MD, PhD (Kyushu University, Fukuoka, Japan); Hiroshi Sugimori, MD, PhD (National Hospital Organization Kyushu Medical Center, Fukuoka, Japan); Shuji Arakawa, MD, PhD (Steel Memorial Yawata Hospital, Kitakyushu, Japan); Kenji Fukuda, MD, PhD (St Mary's Hospital, Kurume, Japan); Tetsuro Ago, MD, PhD (Kyushu University, Fukuoka, Japan); Jiro Kitayama, MD, PhD (Fukuoka Red Cross Hospital, Fukuoka, Japan); Shigeru Fujimoto, MD, PhD (Jichi Medical University, Shimotsuke, Japan); Shoji Arihiro, MD (Japan Labor Health and Welfare Organization Kyushu Rosai Hospital, Kitakyushu, Japan); Junya Kuroda, MD, PhD (National Hospital Organization Fukuoka-Higashi Medical Center, Koga, Japan); Yoshinobu Wakisaka, MD, PhD (Kyushu University Hospital, Fukuoka, Japan); Yoshihisa Fukushima, MD (St Mary's Hospital, Kurume, Japan); Ryu Matsuo, MD, PhD (Secretariat, Kyushu University, Fukuoka, Japan); Fumi Irie, MD, PhD (Kyushu University, Fukuoka, Japan); Kuniyuki Nakamura, MD, PhD (Kyushu University Hospital, Fukuoka, Japan); and Takuya Kiyohara, MD, PhD (Kyushu University Hospital, Fukuoka, Japan).

Conflicts of Interest

None declared.

Multimedia Appendix 1

Stroke prognostic scores.

[[DOCX File, 36 KB - ai_v3i1e46840_app1.docx](#)]

Multimedia Appendix 2

Study populations and events based on the criteria of stroke prognostic scores.

[[DOCX File, 34 KB - ai_v3i1e46840_app2.docx](#)]

Multimedia Appendix 3

Rates of missing values.

[\[DOCX File, 38 KB - ai_v3i1e46840_app3.docx\]](#)

Multimedia Appendix 4

R programs for the development of machine learning-based models.

[\[DOCX File, 33 KB - ai_v3i1e46840_app4.docx\]](#)

Multimedia Appendix 5

Baseline data according to death within 3 months.

[\[DOCX File, 40 KB - ai_v3i1e46840_app5.docx\]](#)**References**

1. Jauch EC, Saver JL, Adams Jr HP, Bruno A, Connors JJ, Demaerschalk BM, et al. Guidelines for the early management of patients with acute ischemic stroke: a guideline for healthcare professionals from the American Heart Association/American Stroke Association. *Stroke* 2013 Mar;44(3):870-947. [doi: [10.1161/STR.0b013e318284056a](#)] [Medline: [23370205](#)]
2. Powers WJ, Rabinstein AA, Ackerson T, Adeoye OM, Bambakidis NC, Becker K, et al. 2018 guidelines for the early management of patients with acute ischemic stroke: a guideline for healthcare professionals from the American Heart Association/American Stroke Association. *Stroke* 2018 Mar;49(3):e46-110 [FREE Full text] [doi: [10.1161/STR.000000000000158](#)] [Medline: [29367334](#)]
3. Hallevi H, Barreto AD, Liebeskind DS, Morales MM, Martin-Schild SB, Abraham AT, et al. Identifying patients at high risk for poor outcome after intra-arterial therapy for acute ischemic stroke. *Stroke* 2009 May;40(5):1780-1785 [FREE Full text] [doi: [10.1161/STROKEAHA.108.535146](#)] [Medline: [19359652](#)]
4. Saposnik G, Kapral MK, Liu Y, Hall R, O'Donnell M, Raptis S, et al. IScore: a risk score to predict death early after hospitalization for an acute ischemic stroke. *Circulation* 2011 Feb 22;123(7):739-749. [doi: [10.1161/CIRCULATIONAHA.110.983353](#)] [Medline: [21300951](#)]
5. Saposnik G, Raptis S, Kapral MK, Liu Y, Tu JV, Mamdani M, et al. The iScore predicts poor functional outcomes early after hospitalization for an acute ischemic stroke. *Stroke* 2011 Dec;42(12):3421-3428. [doi: [10.1161/STROKEAHA.111.623116](#)] [Medline: [21960583](#)]
6. Ntaios G, Faouzi M, Ferrari J, Lang W, Vemmos K, Michel P. An integer-based score to predict functional outcome in acute ischemic stroke: the ASTRAL score. *Neurology* 2012 Jun 12;78(24):1916-1922. [doi: [10.1212/WNL.0b013e318259e221](#)] [Medline: [22649218](#)]
7. O'Donnell MJ, Fang J, D'Uva C, Saposnik G, Gould L, McGrath E, et al. The PLAN score: a bedside prediction rule for death and severe disability following acute ischemic stroke. *Arch Intern Med* 2012 Nov 12;172(20):1548-1556. [doi: [10.1001/2013.jamainternmed.30](#)] [Medline: [23147454](#)]
8. Flint AC, Xiang B, Gupta R, Nogueira RG, Lutsep HL, Jovin TG, et al. THRIVE score predicts outcomes with a third-generation endovascular stroke treatment device in the TREVO-2 trial. *Stroke* 2013 Dec;44(12):3370-3375 [FREE Full text] [doi: [10.1161/STROKEAHA.113.002796](#)] [Medline: [24072003](#)]
9. Gao MM, Wang J, Saposnik G. The art and science of stroke outcome prognostication. *Stroke* 2020 May;51(5):1358-1360. [doi: [10.1161/STROKEAHA.120.028980](#)] [Medline: [32208841](#)]
10. Poldrack RA, Gorgolewski KJ. Making big data open: data sharing in neuroimaging. *Nat Neurosci* 2014 Nov;17(11):1510-1517. [doi: [10.1038/nn.3818](#)] [Medline: [25349916](#)]
11. Shah ND, Steyerberg EW, Kent DM. Big data and predictive analytics: recalibrating expectations. *JAMA* 2018 Jul 03;320(1):27-28. [doi: [10.1001/jama.2018.5602](#)] [Medline: [29813156](#)]
12. Dreiseitl S, Ohno-Machado L. Logistic regression and artificial neural network classification models: a methodology review. *J Biomed Inform* 2002 Oct;35(5-6):352-359 [FREE Full text] [doi: [10.1016/s1532-0464\(03\)00034-0](#)] [Medline: [12968784](#)]
13. Obermeyer Z, Emanuel EJ. Predicting the future - big data, machine learning, and clinical medicine. *N Engl J Med* 2016 Sep 29;375(13):1216-1219 [FREE Full text] [doi: [10.1056/NEJMp1606181](#)] [Medline: [27682033](#)]
14. Handelman GS, Kok HK, Chandra RV, Razavi AH, Lee MJ, Asadi H. eDoctor: machine learning and the future of medicine. *J Intern Med* 2018 Dec;284(6):603-619 [FREE Full text] [doi: [10.1111/joim.12822](#)] [Medline: [30102808](#)]
15. Handelman GS, Kok HK, Chandra RV, Razavi AH, Huang S, Brooks M, et al. Peering into the black box of artificial intelligence: evaluation metrics of machine learning methods. *AJR Am J Roentgenol* 2019 Jan;212(1):38-43. [doi: [10.2214/AJR.18.20224](#)] [Medline: [30332290](#)]
16. Ngiam KY, Khor IW. Big data and machine learning algorithms for health-care delivery. *Lancet Oncol* 2019 May;20(5):e262-e273. [doi: [10.1016/S1470-2045\(19\)30149-4](#)] [Medline: [31044724](#)]
17. Asadi H, Dowling R, Yan B, Mitchell P. Machine learning for outcome prediction of acute ischemic stroke post intra-arterial therapy. *PLoS One* 2014 Feb 10;9(2):e88225 [FREE Full text] [doi: [10.1371/journal.pone.0088225](#)] [Medline: [24520356](#)]

18. van Os HJ, Ramos LA, Hilbert A, van Leeuwen M, van Walderveen MA, Kruyt ND, et al. Predicting outcome of endovascular treatment for acute ischemic stroke: potential value of machine learning algorithms. *Front Neurol* 2018 Sep 25;9:784 [FREE Full text] [doi: [10.3389/fneur.2018.00784](https://doi.org/10.3389/fneur.2018.00784)] [Medline: [30319525](https://pubmed.ncbi.nlm.nih.gov/30319525/)]
19. Heo J, Yoon JG, Park H, Kim YD, Nam HS, Heo JH. Machine learning-based model for prediction of outcomes in acute stroke. *Stroke* 2019 May;50(5):1263-1265. [doi: [10.1161/STROKEAHA.118.024293](https://doi.org/10.1161/STROKEAHA.118.024293)] [Medline: [30890116](https://pubmed.ncbi.nlm.nih.gov/30890116/)]
20. Nishi H, Oishi N, Ishii A, Ono I, Ogura T, Sunohara T, et al. Predicting clinical outcomes of large vessel occlusion before mechanical thrombectomy using machine learning. *Stroke* 2019 Sep;50(9):2379-2388 [FREE Full text] [doi: [10.1161/STROKEAHA.119.025411](https://doi.org/10.1161/STROKEAHA.119.025411)] [Medline: [31409267](https://pubmed.ncbi.nlm.nih.gov/31409267/)]
21. Alaka SA, Menon BK, Brobbey A, Williamson T, Goyal M, Demchuk AM, et al. Functional outcome prediction in ischemic stroke: a comparison of machine learning algorithms and regression models. *Front Neurol* 2020;11:889 [FREE Full text] [doi: [10.3389/fneur.2020.00889](https://doi.org/10.3389/fneur.2020.00889)] [Medline: [32982920](https://pubmed.ncbi.nlm.nih.gov/32982920/)]
22. Brugnara G, Neuberger U, Mahmutoglu MA, Foltyn M, Herweh C, Nagel S, et al. Multimodal predictive modeling of endovascular treatment outcome for acute ischemic stroke using machine-learning. *Stroke* 2020 Dec;51(12):3541-3551. [doi: [10.1161/STROKEAHA.120.030287](https://doi.org/10.1161/STROKEAHA.120.030287)] [Medline: [33040701](https://pubmed.ncbi.nlm.nih.gov/33040701/)]
23. Matsumoto K, Nohara Y, Soejima H, Yonehara T, Nakashima N, Kamouchi M. Stroke prognostic scores and data-driven prediction of clinical outcomes after acute ischemic stroke. *Stroke* 2020 May;51(5):1477-1483. [doi: [10.1161/STROKEAHA.119.027300](https://doi.org/10.1161/STROKEAHA.119.027300)] [Medline: [32208843](https://pubmed.ncbi.nlm.nih.gov/32208843/)]
24. Sirsat MS, Fermé E, Câmara J. Machine learning for brain stroke: a review. *J Stroke Cerebrovasc Dis* 2020 Oct;29(10):105162. [doi: [10.1016/j.jstrokecerebrovasdis.2020.105162](https://doi.org/10.1016/j.jstrokecerebrovasdis.2020.105162)] [Medline: [32912543](https://pubmed.ncbi.nlm.nih.gov/32912543/)]
25. Christodoulou E, Ma J, Collins GS, Steyerberg EW, Verbakel JY, Van Calster B. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *J Clin Epidemiol* 2019 Jun;110:12-22. [doi: [10.1016/j.jclinepi.2019.02.004](https://doi.org/10.1016/j.jclinepi.2019.02.004)] [Medline: [30763612](https://pubmed.ncbi.nlm.nih.gov/30763612/)]
26. Gravesteyn BY, Nieboer D, Ercole A, Lingsma HF, Nelson D, van Calster B, et al. Machine learning algorithms performed no better than regression models for prognostication in traumatic brain injury. *J Clin Epidemiol* 2020 Jun;122:95-107 [FREE Full text] [doi: [10.1016/j.jclinepi.2020.03.005](https://doi.org/10.1016/j.jclinepi.2020.03.005)] [Medline: [32201256](https://pubmed.ncbi.nlm.nih.gov/32201256/)]
27. Cowling TE, Cromwell DA, Bellot A, Sharples LD, van der Meulen J. Logistic regression and machine learning predicted patient mortality from large sets of diagnosis codes comparably. *J Clin Epidemiol* 2021 May;133:43-52. [doi: [10.1016/j.jclinepi.2020.12.018](https://doi.org/10.1016/j.jclinepi.2020.12.018)] [Medline: [33359319](https://pubmed.ncbi.nlm.nih.gov/33359319/)]
28. Kamouchi M, Matsuki T, Hata J, Kuwashiro T, Ago T, Sambongi Y, et al. Prestroke glycemic control is associated with the functional outcome in acute ischemic stroke: the Fukuoka Stroke Registry. *Stroke* 2011 Oct;42(10):2788-2794 [FREE Full text] [doi: [10.1161/STROKEAHA.111.617415](https://doi.org/10.1161/STROKEAHA.111.617415)] [Medline: [21817134](https://pubmed.ncbi.nlm.nih.gov/21817134/)]
29. Kumai Y, Kamouchi M, Hata J, Ago T, Kitayama J, Nakane H, et al. Proteinuria and clinical outcomes after ischemic stroke. *Neurology* 2012 Jun 12;78(24):1909-1915. [doi: [10.1212/WNL.0b013e318259e110](https://doi.org/10.1212/WNL.0b013e318259e110)] [Medline: [22592359](https://pubmed.ncbi.nlm.nih.gov/22592359/)]
30. Quinn TJ, Singh S, Lees KR, Bath PM, Myint PK, VISTA Collaborators. Validating and comparing stroke prognosis scales. *Neurology* 2017 Sep 05;89(10):997-1002 [FREE Full text] [doi: [10.1212/WNL.0000000000004332](https://doi.org/10.1212/WNL.0000000000004332)] [Medline: [28794250](https://pubmed.ncbi.nlm.nih.gov/28794250/)]
31. Hoerl AE, Kennard RW. Ridge regression: biased estimation for nonorthogonal problems. *Technometrics* 1970 Feb;12(1):55-67 [FREE Full text] [doi: [10.1080/00401706.1970.10488634](https://doi.org/10.1080/00401706.1970.10488634)]
32. Tibshirani R. Regression shrinkage and selection via the Lasso. *J R Stat Soc Series B Stat Methodol* 2018 Dec 05;58(1):267-288 [FREE Full text] [doi: [10.1111/j.2517-6161.1996.tb02080.x](https://doi.org/10.1111/j.2517-6161.1996.tb02080.x)]
33. Breiman L. Random forests. *Mach Learn* 2001;45:5-32. [doi: [10.1023/A:1010933404324](https://doi.org/10.1023/A:1010933404324)]
34. Chen TQ, Guestrin C. XGBoost: a scalable tree boosting system. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2016 Aug Presented at: KDD '16; August 13-17, 2016; San Francisco, CA p. 785-794 URL: <https://dl.acm.org/doi/10.1145/2939672.2939785> [doi: [10.1145/2939672.2939785](https://doi.org/10.1145/2939672.2939785)]
35. Ozenne B, Subtil F, Maucort-Boulch D. The precision-recall curve overcame the optimism of the receiver operating characteristic curve in rare diseases. *J Clin Epidemiol* 2015 Aug;68(8):855-859. [doi: [10.1016/j.jclinepi.2015.02.010](https://doi.org/10.1016/j.jclinepi.2015.02.010)] [Medline: [25881487](https://pubmed.ncbi.nlm.nih.gov/25881487/)]
36. Lee SW. Methods for testing statistical differences between groups in medical research: statistical standard and guideline of Life Cycle Committee. *Life Cycle* 2022 Feb 19;2:1-8 [FREE Full text] [doi: [10.54724/lc.2022.e1](https://doi.org/10.54724/lc.2022.e1)]
37. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent Reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *Ann Intern Med* 2015 Jan 06;162(1):55-63 [FREE Full text] [doi: [10.7326/M14-0697](https://doi.org/10.7326/M14-0697)] [Medline: [25560714](https://pubmed.ncbi.nlm.nih.gov/25560714/)]
38. Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, et al. Assessing the performance of prediction models: a framework for traditional and novel measures. *Epidemiology* 2010 Jan;21(1):128-138 [FREE Full text] [doi: [10.1097/EDE.0b013e3181c30fb2](https://doi.org/10.1097/EDE.0b013e3181c30fb2)] [Medline: [20010215](https://pubmed.ncbi.nlm.nih.gov/20010215/)]
39. van der Ploeg T, Austin PC, Steyerberg EW. Modern modelling techniques are data hungry: a simulation study for predicting dichotomous endpoints. *BMC Med Res Methodol* 2014 Dec 22;14:137 [FREE Full text] [doi: [10.1186/1471-2288-14-137](https://doi.org/10.1186/1471-2288-14-137)] [Medline: [25532820](https://pubmed.ncbi.nlm.nih.gov/25532820/)]

40. Steyerberg EW, Uno H, Ioannidis JP, van Calster B, Collaborators. Poor performance of clinical prediction models: the harm of commonly applied methods. *J Clin Epidemiol* 2018 Jun;98:133-143. [doi: [10.1016/j.jclinepi.2017.11.013](https://doi.org/10.1016/j.jclinepi.2017.11.013)] [Medline: [29174118](https://pubmed.ncbi.nlm.nih.gov/29174118/)]
41. Li J, Zhao X, Meng X, Lin J, Liu L, Wang C, et al. High-sensitive C-reactive protein predicts recurrent stroke and poor functional outcome: subanalysis of the clopidogrel in high-risk patients with acute nondisabling cerebrovascular events trial. *Stroke* 2016 Aug;47(8):2025-2030. [doi: [10.1161/STROKEAHA.116.012901](https://doi.org/10.1161/STROKEAHA.116.012901)] [Medline: [27328699](https://pubmed.ncbi.nlm.nih.gov/27328699/)]
42. Hou H, Xiang X, Pan Y, Li H, Meng X, Wang Y. Association of level and increase in D-Dimer with all-cause death and poor functional outcome after ischemic stroke or transient ischemic attack. *J Am Heart Assoc* 2021 Feb 02;10(3):e018600 [FREE Full text] [doi: [10.1161/JAHA.120.018600](https://doi.org/10.1161/JAHA.120.018600)] [Medline: [33412918](https://pubmed.ncbi.nlm.nih.gov/33412918/)]
43. Phipps MS, Cronin CA. Management of acute ischemic stroke. *BMJ* 2020 Feb 13;368:l6983. [doi: [10.1136/bmj.l6983](https://doi.org/10.1136/bmj.l6983)] [Medline: [32054610](https://pubmed.ncbi.nlm.nih.gov/32054610/)]
44. Duncan PW, Bushnell C, Sissine M, Coleman S, Lutz BJ, Johnson AM, et al. Comprehensive stroke care and outcomes: time for a paradigm shift. *Stroke* 2021 Jan;52(1):385-393. [doi: [10.1161/STROKEAHA.120.029678](https://doi.org/10.1161/STROKEAHA.120.029678)] [Medline: [33349012](https://pubmed.ncbi.nlm.nih.gov/33349012/)]

Abbreviations

ASTRAL: Acute Stroke Registry and Analysis of Lausanne

AUPRC: area under the precision-recall curve

AUROC: area under the receiver operating characteristic curve

FSR: Fukuoka Stroke Registry

iScore: Ischemic Stroke Predictive Risk Score

LASSO: least absolute shrinkage and selection operator

PLAN: preadmission comorbidities, level of consciousness, age, and neurologic deficit

RF: random forest

RR: ridge regression

TRIPOD: Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis

XGBoost: Extreme gradient boosting

Edited by K El Emam, B Malin; submitted 27.02.23; peer-reviewed by DK Yon, L Boyer; comments to author 13.09.23; revised version received 30.10.23; accepted 04.12.23; published 11.01.24.

Please cite as:

Irie F, Matsumoto K, Matsuo R, Nohara Y, Wakisaka Y, Ago T, Nakashima N, Kitazono T, Kamouchi M

Predictive Performance of Machine Learning–Based Models for Poststroke Clinical Outcomes in Comparison With Conventional Prognostic Scores: Multicenter, Hospital-Based Observational Study

JMIR AI 2024;3:e46840

URL: <https://ai.jmir.org/2024/1/e46840>

doi: [10.2196/46840](https://doi.org/10.2196/46840)

PMID: [38875590](https://pubmed.ncbi.nlm.nih.gov/38875590/)

©Fumi Irie, Koutarou Matsumoto, Ryu Matsuo, Yasunobu Nohara, Yoshinobu Wakisaka, Tetsuro Ago, Naoki Nakashima, Takanari Kitazono, Masahiro Kamouchi. Originally published in JMIR AI (<https://ai.jmir.org>), 11.01.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR AI, is properly cited. The complete bibliographic information, a link to the original publication on <https://www.ai.jmir.org/>, as well as this copyright and license information must be included.

Original Paper

Improving Risk Prediction of Methicillin-Resistant *Staphylococcus aureus* Using Machine Learning Methods With Network Features: Retrospective Development Study

Methun Kamruzzaman^{1*}, PhD; Jack Heavey^{1*}, BS; Alexander Song^{1*}; Matthew Bielskas^{1*}, MSc; Parantapa Bhattacharya^{1*}, PhD; Gregory Madden^{2*}, MD; Eili Klein^{3,4*}, PhD; Xinwei Deng^{5*}, PhD; Anil Vullikanti^{1,6*}, PhD

¹University of Virginia, Charlottesville, VA, United States

²Division of Infectious Diseases & International Health, Department of Medicine, University of Virginia School of Medicine, Charlottesville, VA, United States

³Department of Emergency Medicine, Johns Hopkins School of Medicine, Baltimore, MD, United States

⁴Center for Disease Dynamics, Economics and Policy, Washington, DC, DC, United States

⁵Department of Statistics, Virginia Tech, Blacksburg, VA, United States

⁶Department of Computer Science, University of Virginia, Charlottesville, VA, United States

* all authors contributed equally

Corresponding Author:

Anil Vullikanti, PhD

University of Virginia

Biocomplexity Institute P.O. Box 400298

Charlottesville, VA, 22904

United States

Phone: 1 5405773102

Email: vsakumar@virginia.edu

Abstract

Background: Health care-associated infections due to multidrug-resistant organisms (MDROs), such as methicillin-resistant *Staphylococcus aureus* (MRSA) and *Clostridioides difficile* (CDI), place a significant burden on our health care infrastructure.

Objective: Screening for MDROs is an important mechanism for preventing spread but is resource intensive. The objective of this study was to develop automated tools that can predict colonization or infection risk using electronic health record (EHR) data, provide useful information to aid infection control, and guide empiric antibiotic coverage.

Methods: We retrospectively developed a machine learning model to detect MRSA colonization and infection in undifferentiated patients at the time of sample collection from hospitalized patients at the University of Virginia Hospital. We used clinical and nonclinical features derived from on-admission and throughout-stay information from the patient's EHR data to build the model. In addition, we used a class of features derived from contact networks in EHR data; these network features can capture patients' contacts with providers and other patients, improving model interpretability and accuracy for predicting the outcome of surveillance tests for MRSA. Finally, we explored heterogeneous models for different patient subpopulations, for example, those admitted to an intensive care unit or emergency department or those with specific testing histories, which perform better.

Results: We found that the penalized logistic regression performs better than other methods, and this model's performance measured in terms of its receiver operating characteristics-area under the curve score improves by nearly 11% when we use polynomial (second-degree) transformation of the features. Some significant features in predicting MDRO risk include antibiotic use, surgery, use of devices, dialysis, patient's comorbidity conditions, and network features. Among these, network features add the most value and improve the model's performance by at least 15%. The penalized logistic regression model with the same transformation of features also performs better than other models for specific patient subpopulations.

Conclusions: Our study shows that MRSA risk prediction can be conducted quite effectively by machine learning methods using clinical and nonclinical features derived from EHR data. Network features are the most predictive and provide significant improvement over prior methods. Furthermore, heterogeneous prediction models for different patient subpopulations enhance the model's performance.

KEYWORDS

methicillin-resistant *Staphylococcus aureus*; network; machine learning; penalized logistic regression; ensemble learning; gradient-boosted classifier; random forest classifier; extreme boosted gradient boosted classifier; Shapley Additive Explanations; SHAP; health care-associated infection; HAI

Introduction

Multidrug-resistant organisms (MDROs), such as *Clostridioides difficile* (CDI), multidrug-resistant gram-negative bacteria (carbapenem-resistant *Acinetobacter baumannii* and carbapenem-resistant Enterobacterales), methicillin-resistant *Staphylococcus aureus* (MRSA), and vancomycin-resistant enterococci, are among the top 10 threats to global health [1]. Health care-associated infections (HAIs) due to MDROs are associated with increased complications, longer hospital stays, and increased mortality. For example, Weiner-Lastinger et al [2] report that HAIs have resulted in billions of dollars in increased healthcare costs [3]. MRSA is one of the most common causes of HAIs and a serious antimicrobial resistance threat, responsible for >10,000 deaths a year in the United States alone [4]. Similar to many other MDROs, MRSA can be easily spread in a hospital from hospitalized patients via contact with the health care environment (ie, shared patient rooms) and health care workers.

Antimicrobial stewardship, which seeks to optimize antibiotic treatment regimens, and infection prevention and control, which involves monitoring, investigating, and managing factors related to MDRO transmission, are the main tools for mitigating the risks of acquisition and severe outcomes of MDROs [5]. Surveillance testing is a critical component of both antimicrobial stewardship and infection prevention control. However, testing is expensive and slow; current laboratory procedures typically require at least 72 hours to report MRSA found in a patient's culture [6]. The delay in testing results in three problems in the hospital: (1) colonized patients remain undetected, leading to potential spread; (2) clinicians treat infections empirically; and (3) increased resource use for contact precautions, leading to both over- and undertreatment.

While several different studies have examined MRSA risk prediction (eg, [6-13]), none to date have progressed to clinical practice due to limitations in generalizability, sample size, and imbalanced data (these are discussed further in the Discussion section). In this study, we demonstrate how improving the hospital context, particularly how patients are connected, can improve the performance of machine learning methods for predicting the outcomes of MRSA surveillance tests, using a rich set of clinical and nonclinical features derived from on-admission and throughout-stay information from a large electronic health record (EHR) data set for patients admitted to the University of Virginia (UVA) Hospital.

Methods

Data Set


We used patient data from the UVA Hospital during 2010-2022. Overall, 27,612 patients in the dataset were tested for MRSA,

and 4171 (15.11%) of them were positive; these patients had 37,237 hospital encounters. The data of each patient's visit can be separated into two parts: (1) on-admission data and (2) clinical event or throughout-stay data, which we have described here:

On-admission data consist of patient demographics and visit information. Patient demographics include information about age, gender, race, ethnicity, country, and state. Visit information includes admission and discharge dates, admission source, admission type, and discharge destination.

Clinical event data represent information collected during the visit. We considered the following event data:

- Procedure: it includes the following kinds of events during this visit or at any time 90 days before this visit: (1) surgeries, (2) device implant or replacement, and (3) dialysis. For a visit, no data after the test collection are used.
- Medication: as MRSA is resistant to specific antibiotics, we also examined prior antibiotic use. We computed the *Days on Therapy*, which indicates whether a patient takes any antibiotic on any specific day. This feature also calculates whether a patient took any antibiotic in the last 90 days of this hospital visit.
- Comorbidity: the International Classification of Diseases, Tenth Revision, code of a patient, which is collected from that patient's medical history, is used to pull comorbidity information using the comorbidity package in R programming language (R Foundation for Statistical Computing). Both Charlson and Elixhauser scores are pulled. It involves other physical conditions such as diabetes, a history of stroke, and a history of dementia.
- MRSA laboratory test: we included both (1) clinical cultures and blood, respiratory, and urine samples collected as part of routine care, which typically requires 48 to 72 hours to return results, and (2) polymerase chain reaction (PCR) surveillance tests, which are administered to MRSA-negative patients admitted to an intensive care unit (ICU; per current hospital policy) or per physician request and typically return results in <72 hours. While surveillance tests provide positive and negative results, clinical cultures may be sent from specimens that are not expected to yield MRSA, even in the presence of an active MRSA infection; therefore, a negative clinical culture result is not considered a definite indicator of noninfection. The nares MRSA PCR likely has equal or higher sensitivity than the nares culture for MRSA [14]. We noted that, in general, testing is not completely unbiased (a patient with an MRSA-positive result admitted to an ICU would not technically need to be screened if they are already on precautions), which might impact the quality of the data set and the results, as we discuss later in the Discussion section.

We applied state-of-the-art machine learning methods to predict the risk of MRSA infection at a given time for a patient, modeled by the outcome of a surveillance test. The data set is split into training (80%) and testing (20%) portions. The model is estimated using the training data, and the hyperparameters are chosen by cross-validation. There are many metrics to evaluate model performance. We used receiver operating characteristics-area under the curve (ROC-AUC) as the overall performance metric of the model (the model evaluation metrics are described in [Multimedia Appendix 1](#)), and a higher value is better. For clinicians, an important objective is to reduce the number of false-negative cases. Therefore, we also used the *false negative rate*  to evaluate the model performance, with a lower value indicating a lower false-negative prediction. The overall model performance is proportional to the ROC-AUC score and inversely proportional to the FNR score.

Problem Statement

The d -days ahead model's MRSA test prediction problem: using features defined from the patient EHR data till some time ($t' = t - d$) predict the outcome of an MRSA surveillance test performed at time t . Formally, let $x(t')$ denote a feature vector for a patient defined till time t and let $y(t)$ denote the result of an MRSA surveillance test performed at time t . The objective is to predict if $y(t) = 1$ using $x(t')$.

The specific questions we study are as follows:

1. How well can MRSA surveillance test results be predicted? What machine learning methods perform well, and what features are the most predictive?
2. Are better predictions possible for specific, meaningful subpopulations?
3. How does the performance vary with d ?
4. Does training with a biased data set (as performed in previous work) impact the true performance?

Interesting Features

Several risk factors for MRSA have been identified in previous studies [15,16]: (1) hospitalization within the past 6 to 12 months, (2) residing in a chronic care facility, (3) being a health care worker, (5) being an intravenous drug user, (5) frequent antibiotic use, (6) antimicrobial therapy within 1 year, (7) history of endotracheal intubation, (8) underlying chronic disorder, (9) presence of an indwelling venous or urinary catheter, (10) history of any surgical procedure, (11) household contact with an identified risk factor, and (12) hypoalbuminemia. We extracted all the aforementioned features from the UVA data set. We created patient-patient and patient-provider interaction networks and extracted the following features from those networks. In addition, we derived many features based on the existing features described in the subsequent section. The total number of features is 108, and the MRSA test outcome is the target feature.

1. Network features: we constructed a contact network $G = (V, E)$ (as shown in [Figure 1](#)), in which we have patient nodes $u_p \in V$ for each patient p and a provider node $u_h \in V$ for each provider h . An edge or contact $(u_{p1}, u_{p2}) \in E$ between 2 patient nodes u_{p1} and u_{p2} indicates that both patients p_1 and p_2 ,

respectively, were colocated (share a common space, a hospital unit in our case) for at least a certain period, in this case at least 900 seconds. Similarly, we defined patient-provider contacts. For instance, in [Figure 1](#), patient P_1 and provider H_1 are colocated at time t_1 , which is represented as edge (u_{p1}, u_{h1}) . The #provider incidents on patient P_1 in the time interval $[t_1, t_2]$ is 2, whereas in the time interval $[t_1, t_3]$, it is 3. We did not use the number of patients and providers that a patient comes into direct contact with as a feature. Instead, we defined slightly different features based on contacts during a time interval, which we found to be more predictive. We take time to be in days. On the basis of the number of contacts for a patient p or a provider h over a period, we constructed the following features:

- **MRSA α** : for a patient p , $S_{p,t}(\alpha) = \{p': (u_p, u_{p'}) \in E, p' \text{ is labeled positive at time } t' \in [t - \alpha, t]\}$, denotes the set of patients who came in contact with p and tested positive in the last α days. We refer to $|S_{p,t}(\alpha)|$ as MRSA α .
- **Provider β** : for a patient p , $\mathcal{S}_{p,t}(\beta) = \{h: (u_p, u_h) \in E, h \text{ visited } p \text{ at time } t' \in [t - \beta, t]\}$. We refer to $|\mathcal{S}_{p,t}(\beta)|$ as Provider β .
- **MRSA positive patients colocated with the patient l** : at the UVA Hospital, patients with an MRSA-positive result might be “cohorted,” that is, they might share a room because they have similar precautions to improve occupancy. For a patient p , let $f_{p,t}(u, \gamma) = \{p': (u_p, u_{p'}) \in E, p' \text{ is labeled positive at } t' \in [t - \gamma, t] \text{ and is in the hospital unit } u \text{ with } p\}$. We referred to $|f_{p,t}(u, \gamma)|$ as the number of patients with colocated MRSA.
- **Bed reuse Π** : let $\Pi_{p,t}(x) = \{p': (u_p, u_{p'}) \notin E, p' \text{ is labeled positive at time } t' < t \text{ and stayed in the same bed } x\}$. We refer to $|\Pi_{p,t}(x)|$ as the number of times Bed x reuse.

Note that all of the aforementioned features are defined for a particular time, t . Therefore, MRSA α and other features should be indexed by the patient and time. To avoid notational clutter, we omit them here when they are clear from the context. For example, suppose $t_1=1, t_2=2, t_3=3, t_4=4$, and $t_5=5$, as shown in [Figure 1](#). Suppose patient P_2 is tested positive at time 4. Then, for patient P_1 , we would have “MRSA 2” at time $t=5$ equal to 1, but “MRSA 2” at time $t=3$ equals 0. For patient P_2 , Provider 2 at time $t=2$ is 0, but Provider 2 at time $t=3$ is 1.

2. Length of stay: for patients p in a hospital encounter, let t_1 denote the admission time and t denote the MRSA test time. The corresponding length of hospital stay (before the MRSA test) was computed as $t-t_1$. For the d -days ($d \geq 0$) ahead model, we computed the corresponding length of stay (before the MRSA test) as $\max\{t-d-t_1, 0\}$. Note that $t-d-t_1$ could be negative if the patient has not been in the hospital long enough—in this case, we took the length of stay to be 0.

3. From the health care facility is a Boolean feature that indicates whether the patient is admitted to the hospital from either “skilled nursing, intermediate care, or assisted living facility” or “long term acute care hospital.” For the d -days ahead model, the feature is defined to be 0 if $t_1-d < 0$, where t_1 is the admission date, and 1, otherwise.

4. δ days observation: we construct several Boolean features based on events in the last δ days before an MRSA test time. For a patient p in a hospital encounter, let $T(e)$ denote the set of times for a specific event e . We defined Boolean variable $e_{\delta}(t) = \{\exists t_1, t_1 \in T(e), t_1 < t, 0 \leq (t - t_1) \leq \delta\}$. We considered $\delta = 90$ and $e \in \{\text{Surgery, Device implant, Antibiotic, Kidney dialysis}\}$. For the d -days ahead model, the feature is defined by considering $\delta + d$ as the parameter in the aforementioned definition, instead of δ .

5. Department-based features: we constructed the following features associated with room stays:

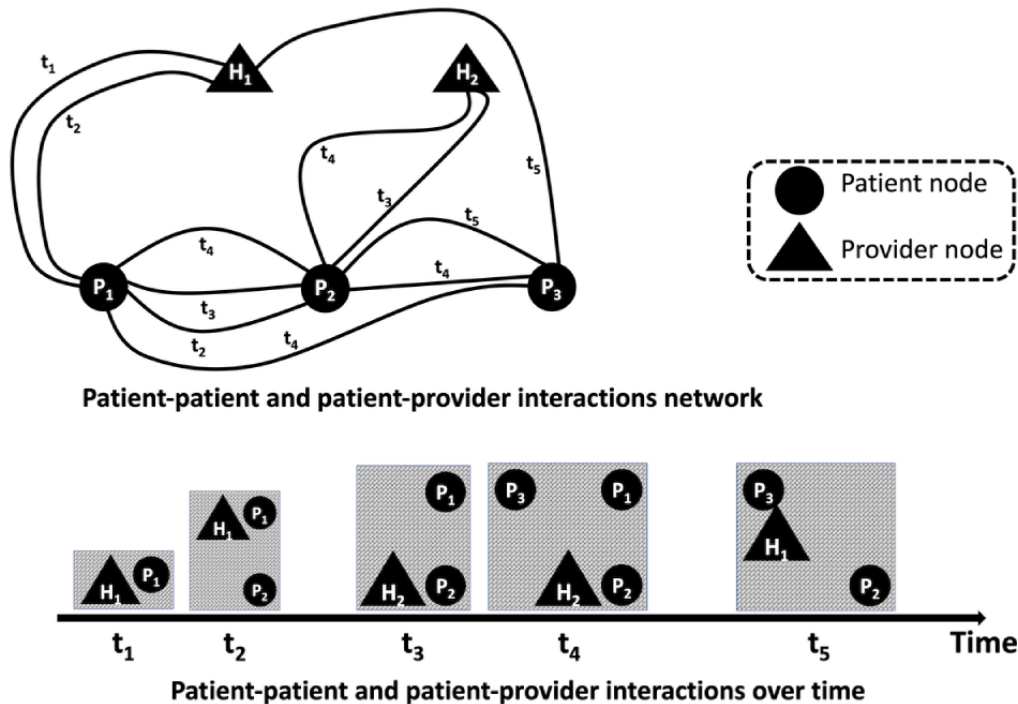
- ICU: this is a Boolean value that indicates whether a patient is admitted to an ICU.
- Emergency department (ED): this is a Boolean value that indicates whether a patient is admitted to the ED.

As in the aforementioned features, for the d -days ahead model, the feature is defined as 1 if the admission to ICU or ED happened before $t - d$, where t is the MRSA test time.

6. PHARMCLASS_k: there are 10 PHARMCLASS (penicillins, miscellaneous anti-infectives, cephalosporins, etc) in the data set. Each PHARMCLASS contains a list of antibiotics. For a patient, PHARMCLASS_k contains the number of antibiotic days from the MRSA testing date in the last 90 days. For the d -days ahead model, the feature is the number of antibiotic days in the 90 days before $t - d$.

7. Test duration days: for a patient p with an MRSA testing date t , we defined this feature as $t - d - t'$, if there exists a time t' , $t(t' < t)$ at which an MRSA test was performed for p ; otherwise, we defined this feature as 0.

Figure 1. Patient-patient and patient-provider interactions are shown on the timeline, where each box represents a room in the hospital, patients are indicated by circles (marked with P) and health care providers are indicated by triangles (marked with H). Multiple patients could share a room, and a provider might visit multiple patients over time. A network is constructed from these interaction events over time. If 2 patients share a room for a certain period (at least for 15 min), we construct an edge between the corresponding patient nodes; similarly, if a provider visits a patient for a certain period (at least for 15 min), we construct an edge between the corresponding patient and provider nodes.



Machine Learning Classifiers

Overview

We explored the following machine learning methods: (1) logistic regression (LR; penalized) [17], (2) support vector machine [18], (3) random forest [19], (4) gradient-boosted classifiers, and (5) XGBoost. These methods have been used extensively on EHR data, and our goal was to understand which ones do well for the MRSA risk-prediction problems we considered in this study. We have described these methods in [Multimedia Appendix 2](#) [17-19]. We also considered these methods with products of features, that is, of the form $x_i(t) \cdot x_j(t)$ where $x_i(t)$ and $x_j(t)$ are different components of the feature vector $x(t)$. We also discuss the Shapley Additive Explanations

(SHAP) technique for understanding feature importance in each model.

Model Explainability Using SHAP

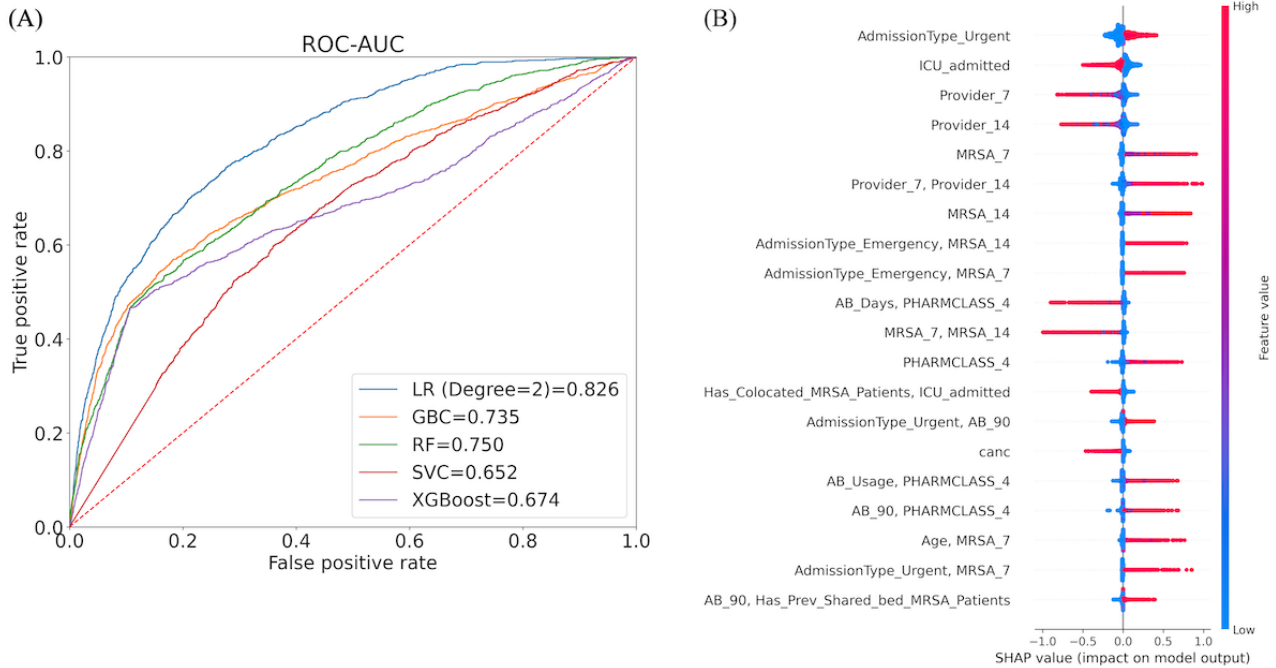
SHAP [20] is a visual feature-attribution process that has many applications in explainable artificial intelligence. It uses a game-theoretic methodology to measure the influence of each feature on the target variable of a machine learning model. Visual representations such as the one in [Figure 2](#), referred to as a summary plot, are used to show the importance of features. The interpretations of this plot are as follows:

- The y-axis specifies the important features arranged from top to bottom regarding their importance (in descending order) to the response variable (the MRSA test result).

- The x-axis indicates the SHAP value of the corresponding feature. The SHAP value of a feature indicates the change in log odds that can be used to extract the probability of success. The color bar on the right-hand side indicates the

- gradient of log odds from low to high, with the color spectrum from blue to red.
- Each point in the SHAP plot for a feature represents an observation of the original data set.

Figure 2. (A) Performance of models on the test data set: performance of different machine learning models on the entire University of Virginia data set. The penalized logistic regression (LR) model with degree-2 features performs best (the receiver operating characteristics-area under the curve [ROC-AUC] for the LR model without feature transformation to degree-2 is 0.734). (B) The most significant features in this model were identified using Shapley Additive Explanations (SHAP). GBC: gradient boosted classifier; RF: random forest; SVC: support vector classifier.



Heterogeneous Risk-Prediction Models for Selected Subpopulations

To improve performance, we developed heterogeneous subpopulation-specific models as described in the subsequent sections.

Based on Testing History

Let $K_{p,t} \in \{+1, -1\}$ denote an MRSA test result for a patient p at time t in a hospital encounter. The testing history $H_{p,t}$ is defined

as $H_{p,t}^j = \{K_{p,t_i} : 1 \leq i \leq j, t_j < t_{j-1} < \dots < t_1 < t\}$. No testing history exists for a newly admitted patient, expressed as $H_{p,t} = \emptyset$. The testing history, considering only the last test result, is expressed as $H_{p,t}^1 = \{K_{p,t}\}$. Similarly, the testing history, considering the last 2 test results, is expressed as $H_{p,t}^2 = \{K_{p,t}, K_{p,t-1}\}$. The number of patients with longer histories drops significantly; therefore, we limited our experiments to the last 2 test results. Table 1 presents the distribution of data points for the different subpopulations.

Table 1. Total number of observations and percentages of positive observations for the subpopulations based on different testing histories.

Previous test history	Total observations	Current test result (-1)	Current test result (+1)	Positive observations
None	27,612	24,371	3241	11.74
-1	11,338	10,179	1159	10.22
+1	3409	863	2546	74.68
(-1, -1)	4755	4320	435	9.15
(-1, +1)	635	198	437	68.82
(+1, -1)	480	328	152	31.67
(+1, +1)	1486	296	1190	80.00

Based on the Admission Source

Recall the Boolean feature named “From health care facility”, which is 1 if the admission source of a patient is a health care

facility. We constructed 2 subpopulations based on whether this feature is 0 or 1; the distributions of these subpopulations and the percentage of positive observations in each are presented in Table 2.

Table 2. Total number of observations and percentages of positive observations for the subpopulations based on different categories.

Subpopulations	Total observations	Test result (-1)	Test result (+1)	Positive observations (%)
Admission source				
Health care facility	2241	1619	622	27.76
Other	42,840	36,198	6642	15.50
Department				
ICU ^a	27,616	24,436	3180	11.52
ED ^b	2538	1658	880	34.67
Other	15,201	11,918	3283	21.60
Hospital stays (days)				
≤15	39,221	32,541	6680	20.53
>15	1643	1413	230	16.28
Antibiotic use (days)				
≤90	30,776	25,065	5711	18.56
>90	16,646	12,997	3649	21.92
0	7097	6368	729	10.27
Age group (years)				
0-50	14,269	12,093	2176	15.25
≥50	27,638	23,008	4630	16.75

^aICU: intensive care unit.

^bED: emergency department.

Based on Department

Recall that both ICU and ED are 2 department-based features, which indicate whether the patient is in the ICU and ED, respectively. The distributions of the subpopulations and the percentage of positive observations are presented in [Table 2](#).

Based on Hospital Stay

The feature “*Length of stay*” captures the number of days a patient has been in the hospital till time $t-d$, where t is the MRSA test date and $d \geq 0$ is the parameter for the d -days ahead model. On the basis of this feature, we constructed 2 subpopulations. The first is the group of patients who have stayed in the hospital for at most 15 days, and the second is the group of patients who have stayed there for >15 days. The distribution of these subpopulations and the percentage of positive observations are presented in [Table 2](#).

Based on Antibiotic Use

Three subpopulations were created based on the number of days for which a patient takes an antibiotic: (1) patients who never took any antibiotics, (2) patients who took antibiotics within the last 90 days from the MRSA testing date, and (3) patients who took antibiotics for more than 90 days from the MRSA testing date. The distribution of these subpopulations and the percentage of positive observations are presented in [Table 2](#).

Based on Age Group

A total of 2 age group-specific patient subgroups, namely 0 to 50 and ≥ 50 years, are considered for the analysis. The

distribution of these subpopulations and the percentage of positive observations are presented in [Table 2](#).

Hierarchical Subpopulation-Based Models

[Figure 3](#) shows the schematic architecture of the hierarchical model. The construction steps of the hierarchical model are as follows:

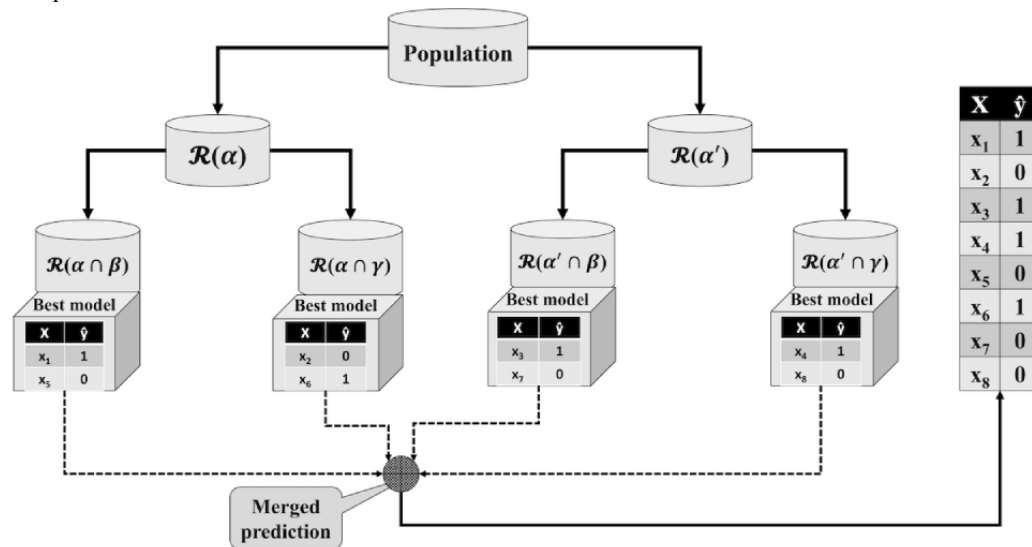
- S1: we defined a set of feature-based rules R at each level to create mutually exclusive subpopulations:
 - At level 1, the rules on the feature named ‘Age-group’ are (1) $R(\alpha)$ =patient subgroup of 0 to 50 years old and (2) $R(\alpha')$ =patient subgroup of more than 50 years old. Each rule creates a patient subpopulation. The patients in these two subpopulations are mutually exclusive, which can be expressed as: $P(\alpha) \cap P(\alpha') = \emptyset$
 - At level 2, each age-group-specific subpopulation is subdivided based on another feature named “Department”. The rules on the ‘Department’ feature are (1) $R(\beta)$ =patient subgroup of ICU and (2) $R(\gamma)$ =patient subgroup of ED. Patients admitted to other departments are not considered in this model.
 - The two-level hierarchical structure creates a set of composite rules (combining rules of each level) at the leaf level that we call two-level rules. The rules are as follows: (a) $R(\alpha \cap \beta)$, (b) $R(\alpha \cap \gamma)$, (c) $R(\alpha' \cap \beta)$, and (d) $R(\alpha' \cap \gamma)$.
- S2: the training population is split based on the 2-level rules. Each training subpopulation is trained on several machine

learning models, and the best-performing model is used for prediction.

- S3: each test observation is passed to the corresponding model using the 2-level rule. The observation with

prediction is stored in a buffer. After completing all the testing observations, the buffer is treated as the model's output.

Figure 3. A schematic view of the hierarchical model architecture. In the figure, X_i represents the i -th observation, y is the model prediction, α is the patient subpopulation who are 0 to 50 years old, α' is the patient subpopulation who are more than 50 years old, β is the patient subpopulation who admitted to intensive care unit (ICU) department, γ is the subpopulation who admitted to the emergency department (ED), and R is a feature-based rule to aggregate data. For instance, $R(\alpha \cap \beta)$ is a 0 to 50 age group patient subpopulation admitted to ICU. At level 1, the overall population is subdivided into two subpopulations based on the feature named "Age-group." The patient subpopulation of age group (0 to 50 years) is mutually exclusive to the patient subpopulation of age group (>50 years). Each age group-specific subpopulation is further subdivided into the next level (level 2) based on another feature named "Department." The patient subpopulation of the ICU department is mutually exclusive to the ED subpopulation. The training data are split based on the 2-level rules, and each patient subpopulation is trained using the best-fitted model. During the testing phase, each data point passes to the appropriate model using the same 2-level rules, and the best-fitted model predicts the outcome. The outcomes of all the models are merged back into the resultant prediction of this hierarchical model.



Data Set for d -Days Ahead Prediction

We prepared a data set to observe the change of prediction performance to the change of d , which is discussed in the Methods section. For each $d \in \{1, 2, \dots, 7\}$, we created a data set, where the feature vector for a patient is generated based on the history of that patient till date $t-d$, where t is the MRSA testing date for that patient.

Ethical Considerations

The data used in the paper was obtained through institutional review board approval and is fully anonymized. Therefore, there are no ethical considerations.

Results

Prediction Model for the Entire Population

We applied multiple machine learning models, including penalized LR, gradient-boosted classifier, Random Forest, support vector classifier, and XGBoost classifier (Multimedia Appendix 2), to the UVA Hospital MRSA patient data sets. We used an 80% to 20% split to construct the train and test data sets. Figure 2A shows the performance of the models. A model's best set of hyperparameters was computed from the training data set using grid search and 10-fold cross-validation. Penalized LR was the best-performing model with the corresponding performance metrics: (1) the FNR score is 0.074, and (2) the

ROC-AUC score is 0.826. Table 3 presents other performance metrics for this data set.

Given the same hyperparameter settings for the penalized LR model, the model performance (ROC-AUC) dropped to 0.734 when we did not consider the product features; therefore, this feature transformation provides a significant benefit. Using the SHAP technique discussed in the Methods section, we extracted the following key features from Figure 2B:

1. "AdmissionType_Urgent," "ICU admitted," "Provider 7," and "Provider 14" are the top 4 features. Recall that "AdmissionType_Urgent" is a Boolean variable where the value 1 indicates the patient admitted as "Urgent." Patients admitted as urgent have a higher likelihood of MRSA infection prediction. Similarly, "ICU admitted" is a Boolean feature where the value 1 indicates that the corresponding patient is admitted to the ICU department and is more likely to predict MRSA infection. On the other hand, "Provider 7" and "Provider 14" indicate the total number of providers a patient contacted in the last 7 and 14 days from the testing date. The higher value of these features is associated with high and negative values for the target feature (MRSA test). A high value comes from the rightmost color bar, and a negative value comes from the x-axis.
2. A high value of "MRSA 7" (which indicates the total number of patients with an MRSA-positive result a patient contacted in the last 7 days from the testing date) is associated with a high and positive value of the target

- feature (the MRSA test); this holds similarly for the “MRSA 14” feature.
- In addition to single features, composite features also correlate more with MRSA infection prediction. For instance, “AdmissionType Emergency” and “MRSA 7” together (similar to “AdmissionType Emergency” and “MRSA 14”) are associated with high and positive values of the target feature (the MRSA test).
 - “PHARMCLASS_4” appears to be an important feature compared to the other PHARMCLASS features. In most cases, this variable is associated with high and positive values for the target feature.

The computational complexity of SHAP increases with the size of the test data set. The best-fitted model is passed to the SHAP explainer method, and it took 5 hours to generate the summary plot (Figure 2B) when the test data set contains 8174 observations and 4656 features. For the same best-fitted model, the SHAP explainer required 1 hour to generate the summary plot when the test data set contained the same number of observations, but the number of features was reduced to 97. Finally, the time was the same when the number of observations in the test data set was reduced to 817, and the number of features was 4656.

Table 3. Performance metrics of the best-performing model for each patient subpopulation based on room allocation, admission source, hospital stay, and antibiotic medication period.

Subpopulation	Model ^a	ROC-AUC ^b	AUPRC ^c	Sensitivity	Specificity	Precision	FPR ^d or fallout	FNR ^e	F ₁ -score	MCC ^f score
Overall	LR ^g	0.826	0.504	0.684	0.797	0.406	0.203	0.074	0.510	0.400
ICU ^h	LR	0.876	0.428	0.775	0.826	0.381	0.174	0.036	0.511	0.455
ED ⁱ	LR	<i>0.936</i> ^j	<i>0.882</i>	<i>0.878</i>	<i>0.886</i>	<i>0.800</i>	<i>0.114</i>	<i>0.067</i>	<i>0.837</i>	<i>0.749</i>
Other rooms	LR	0.752	0.451	0.574	0.793	0.389	0.207	0.110	0.463	0.320
From HCF ^k	LR	0.804	0.585	0.536	0.861	0.571	0.139	0.157	0.553	0.405
Not from HCF	LR	0.831	0.492	0.699	0.801	0.413	0.199	0.070	0.519	0.414
Hospital stay ≤15 days	LR	0.837	0.518	0.722	0.789	0.415	0.211	0.068	0.527	0.421
Hospital stay >15 days	LR	0.729	0.494	0.596	0.803	0.360	0.197	0.086	0.449	0.331
Antibiotic ≤90 days	LR	0.826	0.525	0.681	0.807	0.434	0.193	0.079	0.530	0.416
Antibiotic >90 days	LR	0.841	0.566	0.697	0.809	0.496	0.191	0.092	0.580	0.453
No antibiotic use	LR	0.834	0.328	0.734	0.721	0.201	0.279	<i>0.034</i>	0.315	0.275
Age group (0-50 years)	LR	0.782	0.482	0.613	0.777	0.364	0.223	0.094	0.457	0.325
Age group (≥50 years)	LR	0.833	0.514	0.660	0.817	0.428	0.183	0.079	0.520	0.408
Hierarchical model ^l	HM	<i>0.883</i>	0.490	<i>0.807</i>	0.832	0.440	0.168	<i>0.037</i>	0.569	0.507

^aThis column specifies the best-performing model.

^bROC-AUC: receiver operating characteristics-area under the curve.

^cAUPRC: area under the precision-recall curve.

^dFPR: false positive rate.

^eFNR: false negative rate.

^fMCC: Matthews correlation coefficient.

^gLR: penalized logistic regression.

^hICU: intensive care unit.

ⁱED: emergency department.

^jThe best value for each performance metric is italicized.

^kHCF: health care facility.

^lFor “Hierarchical model” (last row), the highlighted metric (in italics) indicates comparatively better performance than most of the other subpopulations.

Effect of the Imbalanced Data Set

We evaluated the performance achieved using the different sampling techniques discussed earlier. First, as in the study by Hartvigsen et al [8], we used a random selection-based down-sampling technique to select majority-class observations and balance the number of observations between the majority and minority classes. The balanced data are split into train and test data. The ROC-AUC score of the best-performing model on the test data is 0.731. We used the synthetic minority oversampling technique (SMOTE) [21] on our data set to balance both majority and minority classes. The ROC-AUC score of the best-performing model on the test data is 0.896. Similar to the study by Hirano et al [9], we used SMOTE to balance the majority and minority classes in the imbalanced train and test data. The ROC-AUC score of the best-performing model on the test data is 0.903. However, when we evaluated the performance of the abovementioned models on a random test data set, the ROC-AUC score was significantly lower at 0.701. Thus, for our problem, the biased sampling techniques did not improve performance.

Subpopulation-Specific Results

Our models and feature engineering cannot improve the ROC-AUC of 0.826. We now discuss the results of subpopulation-specific models.

Testing History–Based Analysis

The best-fitted model on testing history–based subpopulations (Table 4) showed the best performance on three subpopulations: (1) patients with a (–1) testing history: the best-fitted model had an ROC-AUC of 0.802; (2) patients with a (–1, –1) testing history: the best-fitted model had ROC-AUC of 0.848 and FNR of 0.035; (3) patients with a (+1, +1) testing history: the best model, in terms of the area under the precision-recall curve (AUPRC; Qi et al [22] suggested this metric for imbalanced data) performance metric, had an AUPRC of 0.910 (Figure 4B). The results for the other testing history–based data sets are shown in Multimedia Appendix 3.

Figure 4C shows the significant features (using the SHAP technique) for the (–1, –1) testing history–based subpopulations. The topmost feature (“MRSA 14”) is a network-based feature. Moreover, the network-based features are among the top 10 features. Among these features, “MRSA 7” and “MRSA 14” are positively associated with MRSA infection. In addition to the network features, the interval between the 2 MRSA tests is also important. In addition, patient comorbidity conditions have a significant correlation with MRSA infection.

Table 4. Performance metrics for the best-performing model for each patient subpopulation based on testing history.

Testing history	Model ^a	ROC-AUC ^b	AUPRC ^c	Sensitivity	Specificity	Precision	FPR ^d or fall out	FNR ^e	F ₁ -score	MCC ^f score
None	LR ^g	0.814	0.406	0.689	0.749	0.276	0.251	0.054	0.394	0.311
(–1)	GB ^h	0.802	0.331	0.281	0.953 ⁱ	0.400	0.047	0.078	0.330	0.274
(+1)	LR	0.718	0.884	0.649	0.651	0.847	0.349	0.615	0.735	0.264
(–1, –1)	LR	0.848	0.402	0.697	0.855	0.332	0.145	0.035	0.449	0.404
(–1, +1)	SV ^j	0.613	0.781	0.295	0.897	0.867	0.103	0.639	0.441	0.209
(+1, –1)	SV	0.558	0.614	0.875	0.031	0.311	0.969	0.667	0.459	0.183
(+1, +1)	LR	0.761	0.910	0.595	0.787	0.916	0.213	0.667	0.721	0.308

^aThe “Model” column specifies the best-performing model (LR=penalized logistic regression classifier, GB=gradient boosting, and SV=support vector).

^bROC-AUC: receiver operating characteristics-area under the curve.

^cAUPRC: area under the precision-recall curve.

^dFPR: false positive rate.

^eFNR: false negative rate.

^fMCC: Matthews correlation coefficient.

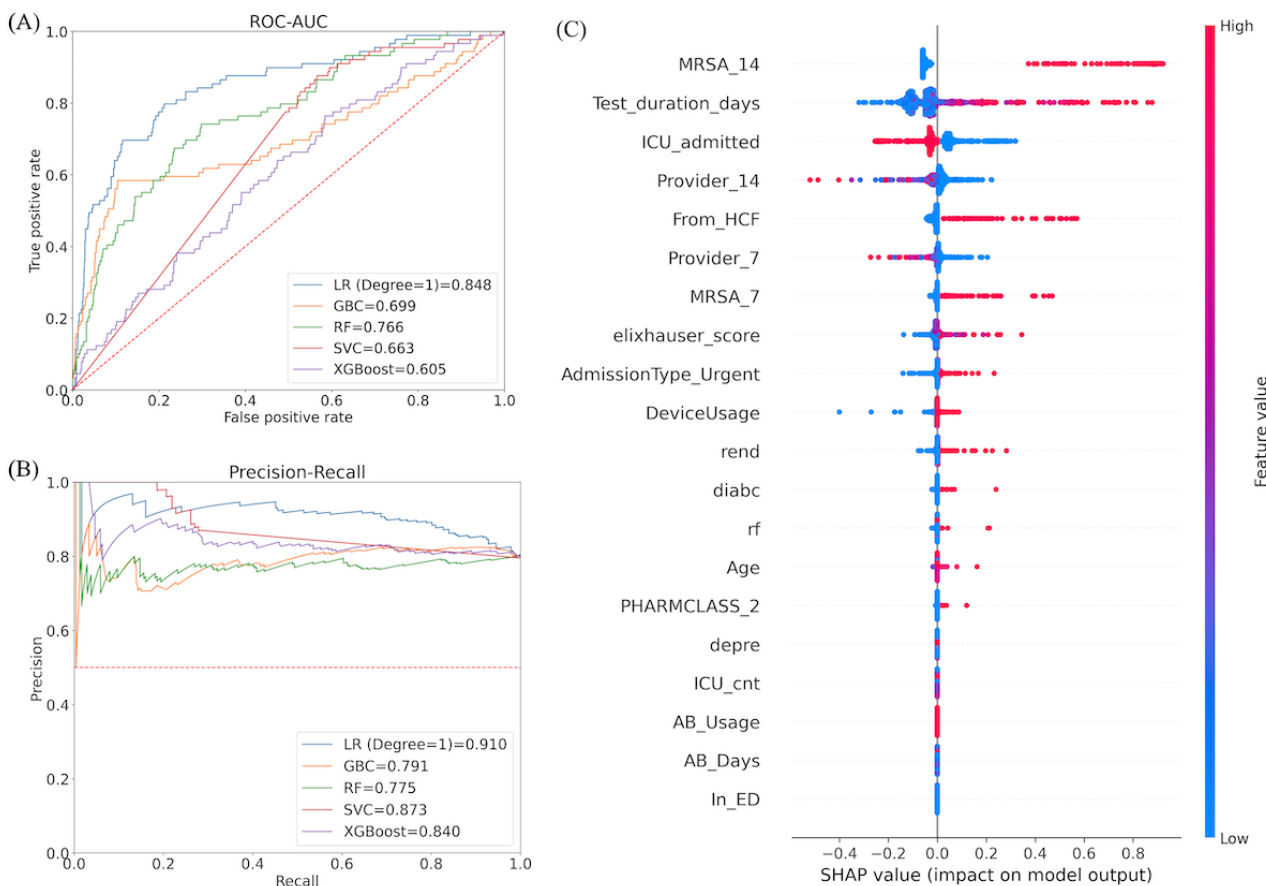
^gLR: logistic regression.

^hGB: gradient boosting.

ⁱThe best value for each performance metric is italicized.

^jSV: support vector.

Figure 4. Results for best-performing subpopulations based on testing history: (A) Performance (receiver operating characteristics-area under the curve [ROC-AUC]) of different machine learning models for testing history (-1, -1), that is, the last 2 testing results are negative—penalized logistic regression (LR) has the best performance. (B) Performance (area under the precision-recall curve [AUPRC]) of different machine learning models for testing history (+1, +1), that is, the last 2 testing results are positive—penalized LR has the best performance. (C) Top features for (-1, -1) testing history-based subpopulation using the LR model. GBC: gradient boosted classifier; RF: random forest; SVC: support vector classifier.



Analysis for ICU and ED Subpopulations

We developed models for other subpopulations, and the performance of the best-fitted models for these subpopulations is reported in Table 3. We found that the best performance is for the ED subpopulation in terms of both ROC-AUC and AUPRC. The ROC-AUC value for the best-fitted model is 0.936 (Figure 5A), and the AUPRC value for the best-fitted model is 0.882 (Figure 5B). Regarding the FNR, the model best performs for the subpopulation without antibiotics. The FNR score obtained using the best-performing model for this data set is 0.034. The subpopulation with the second-best performance is the ICU subpopulation (Figure 6), and the corresponding FNR score is 0.036. The results for the other subpopulations are presented in Multimedia Appendix 4.

Figure 6B shows the significant features (using the SHAP technique) of the best model for the ICU subpopulation. The

top 5 network-based features and the frequency of network features in the top 20 again demonstrate the significance of the network structure. Some of the nonnetwork features that appear to be important are the patient’s age, use of antibiotics in the last 90 days, use of a device in the last 90 days, test duration days, PHARMCLASS 4, and emergency and urgent-type patient admission.

Figure 5C shows the significant features (using the SHAP technique) for the best-performing model for the ED subpopulation. The top 7 features have network features. The top influential feature for the ICU subpopulation is “MRSA 14,” whereas the top significant feature for the ED subpopulation is “MRSA 7.” Unlike in the ICU, the patient’s gender, length of stay, and comorbidity conditions are also crucial in addition to network features.

Figure 5. Results for the emergency department (ED) subpopulation that shows the best performance: (A) performance (receiver operating characteristics-area under the curve [ROC-AUC]) of different machine learning models—penalized logistic regression (LR) has the best performance. (B) Performance (area under the precision-recall curve [AUPRC]) of different machine learning models—penalized LR has the best performance. (C) Top features of the LR model. GBC: gradient boosted classifier; RF: random forest; SHAP: Shapley Additive Explanations; SVC: support vector classifier.

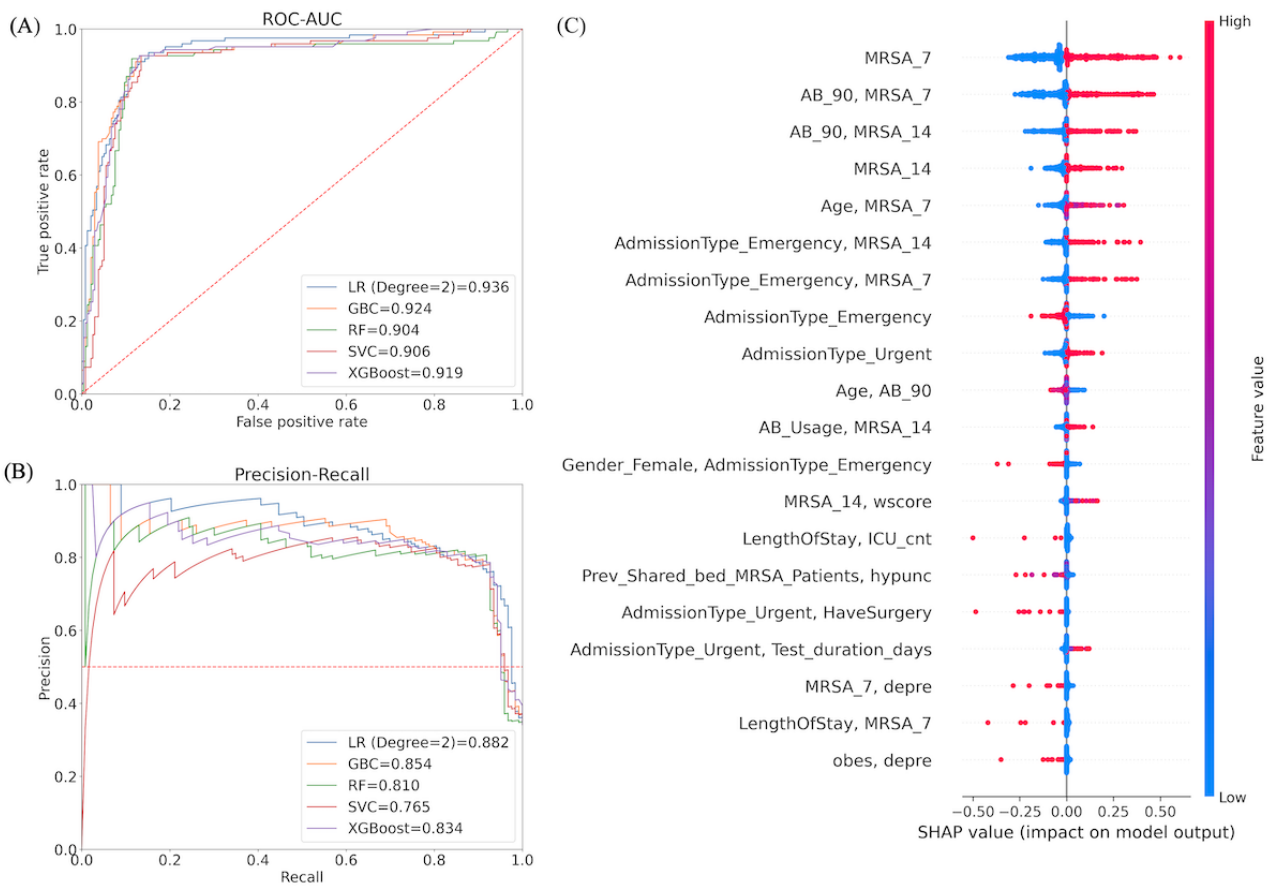
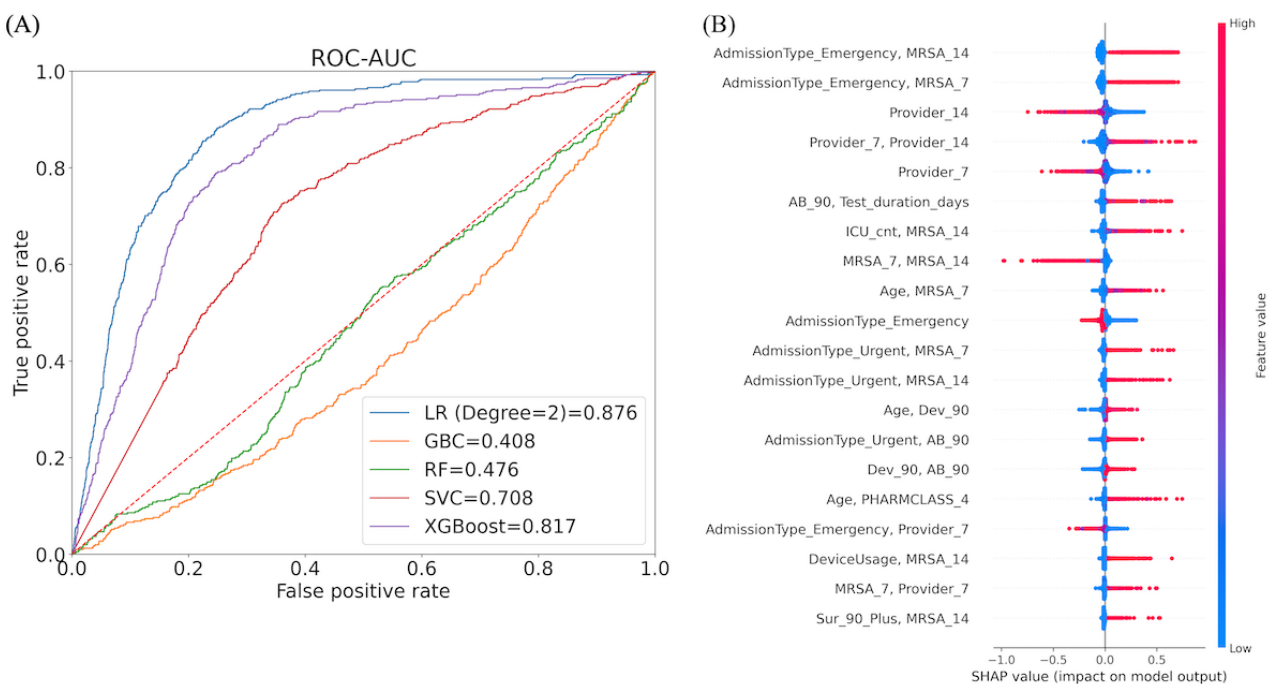


Figure 6. (A) Performance of different machine learning models for the intensive care unit subpopulation; the penalized logistic regression (LR) model performs best. (B) Top features of the LR model. GBC: gradient boosted classifier; RF: random forest; SHAP: Shapley Additive Explanations; SVC: support vector classifier.



Hierarchical Models

The performance of this model is presented in Table 3. This model's ROC-AUC and FNR scores are 0.883 and 0.037, respectively. This model performs better than most subpopulation-based models except for the ED subpopulation-based models.

Importance of Network Features

The best-fitted model performance on the entire data set shows the best performance (Table 3) regarding ROC-AUC and FNR when we use network features. The corresponding ROC-AUC score is 0.826, and the FNR score is 0.074. Without the network features, the ROC-AUC score for the best-fitted model is 0.714, and the FNR score is 0.107 (Table 5).

The ROC-AUC score improved by approximately 16%, and the FNR score improved by approximately 31% because of the network features. The influence of network features is also

significant in the models for the ICU and ED patient subpopulations. The performance metric ROC-AUC improved by approximately 27% for the ICU department patient subpopulation, and the FNR score improved by approximately 58%. For ED patient subpopulations, the performance metric ROC-AUC improved by approximately 30%, the FNR score improved by approximately 69%, and the AUPRC score improved by approximately 50%.

Network features also improve the performance of the best-fitted model for testing history-based subpopulations (Tables 3 and 6).

The ROC-AUC performance metrics for the best-fitted model (−1) testing the history-based subpopulation improved by approximately 11%. For (−1, −1) testing the history-based subpopulation, the best-fitted model performance improved by approximately 25% on the ROC-AUC score and approximately 35% on the FNR score.

Table 5. Performance metrics of the best-performing model for each patient subpopulation based on room allocation, admission source, hospital stay, and antibiotic medication period after excluding the network features.

Subpopulation	Model ^a	AUC ^b	AUPRC ^c	Sensitivity	Specificity	Precision	Fall out	FNR ^d	F ₁ -score	MCC ^e score
Overall	LR ^f	0.714	0.383	0.610	0.709	0.314	0.291	0.107	0.415	0.257
ICU ^g	LR	0.690	0.311	0.547	0.760	0.262	0.240	0.085	0.354	0.233
ED ^h	LR	0.722	0.589	0.593	0.705	0.496	0.295	0.220	0.541	0.287
Other rooms	LR	0.692	0.346	0.631	0.672	0.308	0.328	0.113	0.414	0.243
From HCF ⁱ	LR	0.594	0.340	0.348	0.799	0.375	0.201	0.220	0.361	0.151
Not from HCF	LR	0.721	0.367	0.631	0.704	0.298	0.296	0.095	0.405	0.261
Hospital stay ≤15 days	LR	0.718	0.381	0.615	0.712	0.311	0.288	0.103	0.413	0.261
Hospital stay >15 days	LR	0.595	0.262	0.615	0.566	0.209	0.434	0.112	0.312	0.133
Antibiotic ≤90 days	LR	0.732	0.402	0.634	0.721	0.336	0.279	0.101	0.439	0.288
Antibiotic >90 days	LR	0.707	0.434	0.621	0.683	0.361	0.317	0.138	0.457	0.261
No antibiotic use	LR	0.661	0.236	0.520	0.696	0.178	0.304	0.080 ^j	0.265	0.145
Age group (0-50 years)	LR	0.715	0.404	0.617	0.703	0.298	0.297	0.100	0.402	0.251
Age group (≥50 years)	LR	0.721	0.357	0.628	0.714	0.295	0.286	0.090	0.401	0.265

^aThe "Model" column specifies the best-performing model (LR=penalized logistic regression classifier).

^bAUC: area under the curve.

^cAUPRC: area under the precision-recall curve.

^dFNR: false negative rate.

^eMCC: Matthews correlation coefficient.

^fLR: logistic regression.

^gICU: intensive care unit.

^hED: emergency department.

ⁱHCF: health care facility.

^jitalics.

Table 6. Performance metrics for the best-performing model for each patient subpopulation based on testing history after excluding the network features.

Testing history	Model ^a	AUC ^b	AUPRC ^c	Sensitivity	Specificity	Precision	Fall out	FNR ^d	F ₁ -score	MCC ^e score
None	LR ^f	0.660	0.221	0.565	0.660	0.187	0.340	0.084	0.281	0.153
(-1)	GB ^g	0.723	0.233	0.031	0.996	0.467	0.004	0.098	0.058	0.099
(+1)	LR	0.685	0.851	0.623	0.628	0.821	0.372	0.620	0.708	0.224
(-1, -1)	LR	0.677	0.196	0.663	0.615	0.151	0.385	0.054	0.246	0.164
(-1, +1)	SV ^h	0.637	0.797	0.625	0.615	0.786	0.385	0.579	0.696	0.223
(+1, -1)	SV	0.507	0.356	0.375	0.656	0.353	0.344	0.323	0.364	0.031
(+1, +1)	LR	0.691	0.881	0.605	0.719	0.887	0.281	0.667	0.719	0.267

^aThe “Model” column specifies the best-performing model (LR=penalized logistic regression, GB=gradient boosting, and SV=support vector).

^bAUC: area under the curve.

^cAUPRC: area under the precision-recall curve.

^dFNR: false negative rate.

^eMCC: Matthews correlation coefficient.

^fLR: logistic regression.

^gGB: gradient boosting.

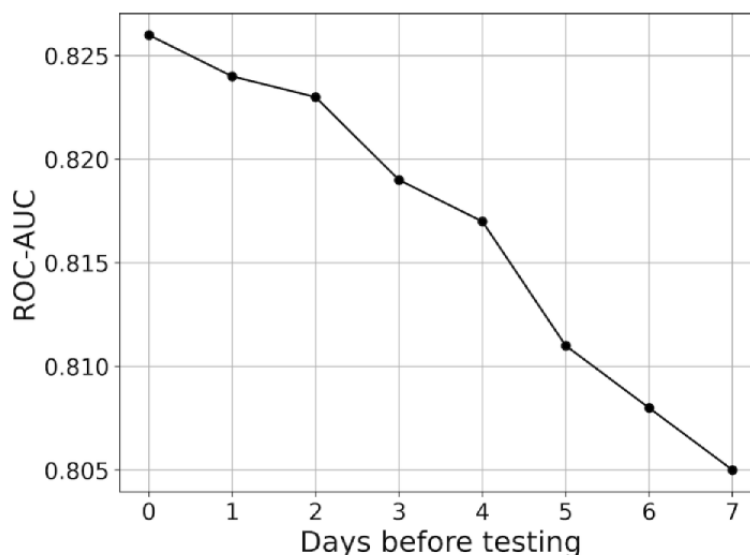
^hSV: support vector.

d-Days Ahead Model Prediction

We now examine how well the test results can be predicted per the d -days ahead model. We expected the performance to drop as d increases, as shown in Figure 7, which shows the

ROC-AUC score of the best-fitted model (for the data set corresponding to d -days before the test, as described in the Methods section) versus d . Note that the performance decays significantly with d .

Figure 7. d -days ahead prediction: performance (receiver operating characteristics-area under the curve [ROC-AUC]) of best model versus d . The performance drops gradually with d .



Discussion

Principal Findings

Our results demonstrate that clinically relevant models can be developed for predicting MRSA test results with high accuracy using a combination of clinical and nonclinical features from EHR data. In particular, features of contact networks (eg, “MRSA 7,” “MRSA 14,” “Provider 7,” and “Provider 14”) constructed from EHR data are quite significant in our models.

Tables 5 and 6 show the performance of the models on the same group of data sets without considering the network features. The empirical results establish that the network features have a significant impact (model performance ROC-AUC improves by > 15%) on MRSA infection prediction.

We took the simplest approach to network construction, which views edges as unweighted, and did not consider heterogeneity in contacts, for example, based on types of providers. It is interesting that even the simplest approach improves performance. While more characteristics of networks and edge

weights could be considered and these might improve the performance, the value of our simple approach is that it is easier to construct and is likely more generalizable and robust because there might be uncertainties in some of these additional characteristics.

In addition to network features, we observed that features associated with antibiotic use (“Antibiotic days”, “Antibiotic days in last 90 days”, “Antibiotic days in last 90+ days”, “PHARMCLASS_1” to “PHARMCLASS_10”, etc.), different kinds of events in the past 90 days (eg, kidney dialysis, device use, and any surgery), and comorbidity conditions such as diabetes without complications (diab or diabunc), hypothyroidism (hypothy), uncomplicated hypertension (hypunc), the Charlson score, the Elixhauser score, the weighted version of the Elixhauser score using the van Walraven algorithm (wscore vw), the weighted version of the Elixhauser score using the *Agency for Healthcare Research and Quality* (AHRQ) algorithm (wscore ahrq), and the weighted version of the Charlson score (wscore) are also predictive; many of these have been identified as important in prior work.

The penalized LR model with degree-2 polynomial features performs best in almost all settings, using a new class of network-based features derived from EHR data. Our results also showed the utility of heterogeneous models for different subpopulations instead of just one model for the entire population. In particular, we obtained good performance for subpopulations in an ICU or ED and those with certain test histories. We also observed that the performance degrades gradually for a d -days ahead prediction.

The testing policy is fairly systematic for patients in the ICU. Therefore, we expect the model for ICU subpopulations to be quite robust and generalizable to data sets from other locations. On the other hand, it is important to note that testing in the entire patient population is generally not completely systematic and might have biases because it is administered per physician request. It is unclear what the impact of these biases would be on the model’s generalizability. A mitigating factor is that the model for the entire population is quite close to that for the ICU, and many of the significant factors are the same. This suggests that the model for the entire population might also be quite robust. Future studies on other data sets are required to determine the generalizability of these models.

Our prediction model for a patient on day t only used features that were available for that patient before day t . This included the network features. Therefore, if a patient was in the hospital for <7 days, the “MRSA 7” and “Provider 7” feature values will be 0, and if a patient was in the hospital for <14 days, the “MRSA 14” and “Provider 14” feature values will be 0. It is possible that the predictive model would be more informative for patients who have a longer history in the hospital, but even this is an important patient population from a clinical perspective.

Finally, we noted that the simple penalized LR model seems to work quite well when given more complex features, such as second-degree features. It is not completely clear why this works much better than the other methods, namely support vector machine, random forest, gradient-boosted classifiers, and

XGBoost. One possible explanation can be because of the model parsimony of the penalized LR. Further research on model validation can be useful. One advantage of our analysis is that the penalized LR method is easy to interpret.

Our models are the most useful for clinical decisions about empiric antibiotic use. For instance, if the test prediction is negative, a clinician could be more comfortable starting an antibiotic treatment. If the test prediction is positive in the context of a newly identified infection, a clinician might consider the benefits of starting an anti-MRSA antibiotic. Isolation precautions are known to have many adverse effects (eg, fewer clinician visits to the room, patient depression, and noninfectious adverse events such as blood clots), although they help in reducing transmission. If the d -days ahead result is negative in a current patient with a positive MRSA result, an epidemiologist may adjust for an earlier test for clearance of isolation precautions.

Comparison With Prior Work

Machine learning using EHR data for clinical informatics is a very active area of research [23,24]. Diverse kinds of statistical and machine learning methods, including deep-learning algorithms, have been used to predict important clinical events (eg, hypertension, diabetes, chronic obstructive pulmonary disease, arrhythmia, asthma, gastritis, dementia, delirium, *Clostridium difficile* infection, and HAIs) using EHR data [8,9,12,13,25-29]. In the context of HAIs, risk-prediction models have been developed for several MDROs. We have briefly discussed examples of such studies to illustrate the types of questions and methods that have been considered, with a focus on MRSA.

Hartvigsen et al [8] and Hirano et al [9] studied a similar problem, namely, predicting MRSA test outcomes, using the Medical Information Mart for Intensive Care III and IV data sets, respectively. These data sets are critical care data sets comprising 12 years (2001 to 2012 and 2008 to 2019, respectively) of patient records from the Beth Israel Deaconess Medical Center Intensive Care Unit in Boston, Massachusetts [11]. Hartvigsen et al [8] show high performance for the prediction of MRSA test outcomes 1 day ahead using subsampled data. Hirano et al [9] achieve high performance (an ROC-AUC value of 0.89) for a slightly different patient subpopulation using the SMOTE [21] technique for handling data imbalance. Rhodes et al [12] consider a slightly different question regarding MRSA infection 72 hours after admission. They show that the Classification Tree Analysis has good performance for the population of patients from the Northwestern Memorial Hospital and Lake Forest Hospital. A review by Tang et al [13] notes that penalized LR, decision tree, and random forest are the preferred methods for antimicrobial resistance prediction.

A significant challenge here all MRSA risk-prediction problems (including our study) is that the data are quite imbalanced because the fraction of positive observations is quite small. Consequently, the performance of most machine learning methods can be affected. A common strategy to address this issue has been to construct data sets using different kinds of sampling techniques, including biased sampling [8,10] and

SMOTE [30]. While this kind of approach can appear to have very good performance on a similarly constructed test data set, the true performance on an unbiased data set might be reduced (as discussed in the study by Pencina et al [31] and in our Results section), which impacts its performance when used in practice. According to the study by Soltanzadeh and Hashemzadeh [30], resolving the class distribution problem using synthetic or biased data constructed in this manner causes many issues such as (1) generalization problems because of noisy samples; (2) uninformative samples; and (3) newly created points being close to the minority class points, which often create points around the decision boundary. Azizi et al [32] and Kokosi and Harron [33] note that (1) the use of synthetic data in the decision-making process and (2) the problem of attribute disclosure are other limitations of using synthetic data.

Our study differs from prior work in 3 ways. First, we used network features in addition to other EHR-based features in our risk-prediction models. It has been shown that network properties are predictive of infection risk, for example, Klein et al [34] showed that patient degree is associated with vancomycin-resistant enterococci risk. Similarly, Riaz et al [35] show that local colonization pressure, which is based on the network structure, is associated with *C. difficile* infection (CDI) risk. Similarly, Miller et al [36] show that household exposure (which can also be viewed as a network effect) increases CDI risk. However, our work is the first to explicitly consider EHR-based features for MRSA test prediction as a machine learning task that can be used in a clinical setting. Second, we identified heterogeneous models for specific patient subgroups and showed that these have significantly better performance. Finally, we developed our prediction models without any biased sampling techniques.

Limitations

We have not been able to improve the ROC-AUC performance of our models above 0.90. Data imbalance and patient diversity could be significant reasons for this performance. As noted

earlier, MRSA infections are fairly rare, and for the problem of MRSA test results, only about 15% of the results are positive. We also note that there are many other notions of MRSA risk, such as the risk of severe outcomes and MRSA acquisition, which we study here. These notions are harder to formalize and learn because the data sets would become even more biased than what we consider here, and new methods are needed for them.

While our results show that network features are the most predictive, there might be uncertainties in inferring them from the EHR data. We note that these (eg, the #providers within a time interval) are not directly available in the patient's EHR data; we are inferring them through colocation information. It is possible that many interactions are not recorded accurately or the times might not be accurate. More work is needed to fully understand the impact of these uncertainties.

Another issue is the testing bias. As discussed earlier, the entire patient population data set has biases because testing is not very systematic in general. This might have an impact on the model's performance when applied to data sets from other hospitals, and the model would have to be retrained. However, the model structure and specific features might still be relevant, especially because they hold for the ICU patient subpopulation, for which testing is more systematic.

Conclusions

Preprocessing by clustering has been useful in many applications. One challenge in using this approach is that a distance metric needs to be defined, which is difficult due to the diversity of features. For instance, some features are datetime related, some are Boolean and categorical, while others are real valued. A possible extension is to transform the features into a latent space, where distances can be computed. Additional feature engineering and more advanced machine learning methods might be useful for further improving performance. In particular, text analysis might be helpful in further improving the performance.

Acknowledgments

This study was partially supported by the Centers for Disease Control and Prevention MInD-Healthcare Program (grant U01CK000589) and NSF grants CCF-1918656 and IIS-1955797. GM is an iTHRIV scholar. The iTHRIV Scholars Program is supported in part by the National Center for Advancing Translational Sciences of the National Institutes of Health under award numbers UL1TR003015 and KL2TR003016.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Machine learning model evaluation metrics.

[PDF File (Adobe PDF File), 174 KB - [ai_v3i1e48067_app1.pdf](#)]

Multimedia Appendix 2

Machine learning models.

[PDF File (Adobe PDF File), 91 KB - [ai_v3i1e48067_app2.pdf](#)]

Multimedia Appendix 3

Test history-based results and machine learning model hyperparameters.

[PDF File (Adobe PDF File), 605 KB - [ai_v3i1e48067_app3.pdf](#)]

Multimedia Appendix 4

Patient subpopulation-based results and machine learning model hyperparameters.

[PDF File (Adobe PDF File), 989 KB - [ai_v3i1e48067_app4.pdf](#)]

References

1. Shallcross LJ, Davies SC. The World Health Assembly resolution on antimicrobial resistance. *J Antimicrob Chemother* 2014 Nov;69(11):2883-2885. [doi: [10.1093/jac/dku346](#)] [Medline: [25204342](#)]
2. Weiner-Lastinger LM, Abner S, Edwards JR, Kallen AJ, Karlsson M, Magill SS, et al. Antimicrobial-resistant pathogens associated with adult healthcare-associated infections: summary of data reported to the National Healthcare Safety Network, 2015-2017. *Infect Control Hosp Epidemiol* 2020 Jan;41(1):1-18 [FREE Full text] [doi: [10.1017/ice.2019.296](#)] [Medline: [31767041](#)]
3. Zimlichman E, Henderson D, Tamir O, Franz C, Song P, Yamin CK, et al. Health care-associated infections: a meta-analysis of costs and financial impact on the US health care system. *JAMA Intern Med* 2013 Dec;173(22):2039-2046. [doi: [10.1001/jamainternmed.2013.9763](#)] [Medline: [23999949](#)]
4. 2019 AR threats report. Centers for Disease Control and Prevention. URL: <https://www.cdc.gov/drugresistance/biggest-threats.html> [accessed 2024-04-04]
5. Core elements of hospital antibiotic stewardship programs. Centers for Disease Control and Prevention. URL: <https://www.cdc.gov/antibiotic-use/core-elements/hospital.html#:~:text=Reporting%3A%20Regularly%20report%20information%20on,an%20resistance%20and%20optimal%20prescribing> [accessed 2024-04-04]
6. Shang JS, Lin YS, Goetz AM. Diagnosis of MRSA with neural networks and logistic regression approach. *Health Care Manag Sci* 2000 Sep;3(4):287-297. [doi: [10.1023/a:1019018129822](#)] [Medline: [11105415](#)]
7. Dutta R, Dutta R. "Maximum probability rule" based classification of MRSA infections in hospital environment: using electronic nose. *Sens Actuators B Chem* 2006 Dec 14;120(1):156-165. [doi: [10.1016/j.snb.2006.02.013](#)]
8. Hartvigsen T, Sen C, Brownell S, Teeple E, Kong X, Rundensteiner E. Early prediction of MRSA infections using electronic health records. In: Proceedings of the 11th International Conference on Health Informatics. 2018 Presented at: HEALTHINF 2018; January 19-21, 2018; Madeira, Portugal. [doi: [10.5220/0006599601560167](#)]
9. Hirano Y, Shinmoto K, Okada Y, Suga K, Bombard J, Murahata S, et al. Machine learning approach to predict positive screening of Methicillin-resistant Staphylococcus aureus during mechanical ventilation using synthetic dataset from MIMIC-IV database. *Front Med (Lausanne)* 2021 Nov 16;8:694520 [FREE Full text] [doi: [10.3389/fmed.2021.694520](#)] [Medline: [34869405](#)]
10. Hsu CC, Lin YE, Chen YS, Liu YC, Muder RR. Validation study of artificial neural network models for prediction of methicillin-resistant Staphylococcus aureus carriage. *Infect Control Hosp Epidemiol* 2008 Jul;29(7):607-614. [doi: [10.1086/588588](#)] [Medline: [18549315](#)]
11. Johnson AE, Pollard TJ, Shen L, Lehman LW, Feng M, Ghassemi M, et al. MIMIC-III, a freely accessible critical care database. *Sci Data* 2016 May 24;3:160035 [FREE Full text] [doi: [10.1038/sdata.2016.35](#)] [Medline: [27219127](#)]
12. Rhodes NJ, Rohani R, Yarnold PR, Pawlowski AE, Malczynski M, Qi C, et al. Machine learning to stratify methicillin-resistant staphylococcus aureus risk among hospitalized patients with community-acquired pneumonia. *Antimicrob Agents Chemother* 2023 Jan 24;67(1):e0102322 [FREE Full text] [doi: [10.1128/aac.01023-22](#)] [Medline: [36472425](#)]
13. Tang R, Luo R, Tang S, Song H, Chen X. Machine learning in predicting antimicrobial resistance: a systematic review and meta-analysis. *Int J Antimicrob Agents* 2022;60(5-6):106684. [doi: [10.1016/j.ijantimicag.2022.106684](#)] [Medline: [36279973](#)]
14. Shenoy ES, Noubary F, Kim J, Rosenberg ES, Cotter JA, Lee H, et al. Concordance of PCR and culture from nasal swabs for detection of methicillin-resistant Staphylococcus aureus in a setting of concurrent antistaphylococcal antibiotics. *J Clin Microbiol* 2014 Apr;52(4):1235-1237 [FREE Full text] [doi: [10.1128/JCM.02972-13](#)] [Medline: [24452168](#)]
15. Boyce JM, Potter-Bynoe G, Chenevert C, King T. Environmental contamination due to methicillin-resistant Staphylococcus aureus: possible infection control implications. *Infect Control Hosp Epidemiol* 1997 Sep;18(9):622-627. [Medline: [9309433](#)]
16. Herold BC, Immergluck LC, Maranan MC, Lauderdale DS, Gaskin RE, Boyle-Vavra S, et al. Community-acquired methicillin-resistant Staphylococcus aureus in children with no identified predisposing risk. *JAMA* 1998 Feb 25;279(8):593-598. [doi: [10.1001/jama.279.8.593](#)] [Medline: [9486753](#)]
17. Nick TG, Campbell KM. Logistic regression. *Methods Mol Biol* 2007;404:273-301. [doi: [10.1007/978-1-59745-530-5_14](#)] [Medline: [18450055](#)]
18. Cortes C, Vapnik V. Support-vector networks. *Mach Learn* 1995 Sep;20:273-297. [doi: [10.1007/bf00994018](#)]
19. Leo B. Random forests. *Mach Learn* 2001;45:5-32 [FREE Full text] [doi: [10.1023/A:1010933404324](#)]
20. Lundberg S, Lee SI. A unified approach to interpreting model predictions. arXiv Preprint posted online May 22, 2017 [FREE Full text]

21. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. *J Artif Intell Res* 2002 Jun 01;16:321-357. [doi: [10.1613/jair.953](https://doi.org/10.1613/jair.953)]
22. Qi Q, Luo Y, Xu Z, Ji S, Yang T. Stochastic optimization of areas under precision-recall curves with provable convergence. *arXiv Preprint* posted online April 18, 2021 [[FREE Full text](#)]
23. Jordan MI, Mitchell TM. Machine learning: trends, perspectives, and prospects. *Science* 2015 Jul 17;349(6245):255-260. [doi: [10.1126/science.aaa8415](https://doi.org/10.1126/science.aaa8415)] [Medline: [26185243](https://pubmed.ncbi.nlm.nih.gov/26185243/)]
24. Wiens J, Shenoy ES. Machine learning for healthcare: on the verge of a major shift in healthcare epidemiology. *Clin Infect Dis* 2018 Jan 06;66(1):149-153 [[FREE Full text](#)] [doi: [10.1093/cid/cix731](https://doi.org/10.1093/cid/cix731)] [Medline: [29020316](https://pubmed.ncbi.nlm.nih.gov/29020316/)]
25. Bhagwat N, Viviano JD, Voineskos AN, Chakravarty MM, Alzheimer's Disease Neuroimaging Initiative. Modeling and prediction of clinical symptom trajectories in Alzheimer's disease using longitudinal data. *PLoS Comput Biol* 2018 Sep 14;14(9):e1006376 [[FREE Full text](#)] [doi: [10.1371/journal.pcbi.1006376](https://doi.org/10.1371/journal.pcbi.1006376)] [Medline: [30216352](https://pubmed.ncbi.nlm.nih.gov/30216352/)]
26. Cleret de Langavant L, Bayen E, Yaffe K. Unsupervised machine learning to identify high likelihood of dementia in population-based surveys: development and validation study. *J Med Internet Res* 2018 Jul 09;20(7):e10493 [[FREE Full text](#)] [doi: [10.2196/10493](https://doi.org/10.2196/10493)] [Medline: [29986849](https://pubmed.ncbi.nlm.nih.gov/29986849/)]
27. Oh J, Makar M, Fusco C, McCaffrey R, Rao K, Ryan EE, et al. A generalizable, data-driven approach to predict daily risk of clostridium difficile infection at two large academic health centers. *Infect Control Hosp Epidemiol* 2018 Apr;39(4):425-433 [[FREE Full text](#)] [doi: [10.1017/ice.2018.16](https://doi.org/10.1017/ice.2018.16)] [Medline: [29576042](https://pubmed.ncbi.nlm.nih.gov/29576042/)]
28. Wong A, Young AT, Liang AS, Gonzales R, Douglas VC, Hadley D. Development and validation of an electronic health record-based machine learning model to estimate delirium risk in newly hospitalized patients without known cognitive impairment. *JAMA Netw Open* 2018 Aug 03;1(4):e181018 [[FREE Full text](#)] [doi: [10.1001/jamanetworkopen.2018.1018](https://doi.org/10.1001/jamanetworkopen.2018.1018)] [Medline: [30646095](https://pubmed.ncbi.nlm.nih.gov/30646095/)]
29. Yang Z, Huang Y, Jiang Y, Sun Y, Zhang YJ, Luo P. Clinical assistant diagnosis for electronic medical record based on convolutional neural network. *Sci Rep* 2018 Apr 20;8(1):6329 [[FREE Full text](#)] [doi: [10.1038/s41598-018-24389-w](https://doi.org/10.1038/s41598-018-24389-w)] [Medline: [29679019](https://pubmed.ncbi.nlm.nih.gov/29679019/)]
30. Soltanzadeh P, Hashemzadeh M. RCSMOTE: range-controlled synthetic minority over-sampling technique for handling the class imbalance problem. *Inf Sci* 2021 Jan 04;542:92-111. [doi: [10.1016/j.ins.2020.07.014](https://doi.org/10.1016/j.ins.2020.07.014)]
31. Pencina MJ, Goldstein BA, D'Agostino RB. Prediction models - development, evaluation, and clinical application. *N Engl J Med* 2020 Apr 23;382(17):1583-1586. [doi: [10.1056/NEJMp2000589](https://doi.org/10.1056/NEJMp2000589)] [Medline: [32320568](https://pubmed.ncbi.nlm.nih.gov/32320568/)]
32. Azizi Z, Zheng C, Mosquera L, Pilote L, El Emam K, GOING-FWD Collaborators. Can synthetic data be a proxy for real clinical trial data? A validation study. *BMJ Open* 2021 Apr 16;11(4):e043497 [[FREE Full text](#)] [doi: [10.1136/bmjopen-2020-043497](https://doi.org/10.1136/bmjopen-2020-043497)] [Medline: [33863713](https://pubmed.ncbi.nlm.nih.gov/33863713/)]
33. Kokosi T, Harron K. Synthetic data in medical research. *BMJ Med* 2022 Sep 26;1(1):e000167 [[FREE Full text](#)] [doi: [10.1136/bmjmed-2022-000167](https://doi.org/10.1136/bmjmed-2022-000167)] [Medline: [36936569](https://pubmed.ncbi.nlm.nih.gov/36936569/)]
34. Klein EY, Tseng KK, Hinson J, Goodman KE, Smith A, Toerper M, et al. The role of healthcare worker-mediated contact networks in the transmission of vancomycin-resistant enterococci. *Open Forum Infect Dis* 2020 Feb 15;7(3):ofaa056 [[FREE Full text](#)] [doi: [10.1093/ofid/ofaa056](https://doi.org/10.1093/ofid/ofaa056)] [Medline: [32166095](https://pubmed.ncbi.nlm.nih.gov/32166095/)]
35. Riaz T, Khan N, Polgreen P, Segre A, Sewell D, Pemmaraju S. Highly local *Clostridioides difficile* infection (CDI) pressure as risk factors for CDI. *Infect Control Hosp Epidemiol* 2020 Nov 02;41(S1):s250. [doi: [10.1017/ice.2020.810](https://doi.org/10.1017/ice.2020.810)]
36. Miller AC, Arakkal AT, Sewell DK, Segre AM, Pemmaraju SV, Polgreen PM. Risk for asymptomatic household transmission of *Clostridioides difficile* infection associated with recently hospitalized family members. *Emerg Infect Dis* 2022 May;28(5):932-939 [[FREE Full text](#)] [doi: [10.3201/eid2805.212023](https://doi.org/10.3201/eid2805.212023)] [Medline: [35447064](https://pubmed.ncbi.nlm.nih.gov/35447064/)]

Abbreviations

- AUPRC:** area under the precision-recall curve
- ED:** emergency department
- EHR:** electronic health record
- FNR:** false negative rate
- HAI:** health care-associated infection
- ICU:** intensive care unit
- LR:** logistic regression
- MDRO:** multidrug-resistant organism
- MRSA:** methicillin-resistant *Staphylococcus aureus*
- ROC-AUC:** receiver operating characteristics-area under the curve
- SHAP:** Shapley Additive Explanations
- SMOTE:** synthetic minority oversampling technique
- UVA:** University of Virginia

Edited by K El Emam, B Malin; submitted 10.04.23; peer-reviewed by D Sewell, B Zhao; comments to author 02.07.23; revised version received 28.09.23; accepted 13.01.24; published 16.05.24.

Please cite as:

*Kamruzzaman M, Heavey J, Song A, Bielskas M, Bhattacharya P, Madden G, Klein E, Deng X, Vullikanti A
Improving Risk Prediction of Methicillin-Resistant Staphylococcus aureus Using Machine Learning Methods With Network Features:
Retrospective Development Study*

JMIR AI 2024;3:e48067

URL: <https://ai.jmir.org/2024/1/e48067>

doi: [10.2196/48067](https://doi.org/10.2196/48067)

PMID: [38875598](https://pubmed.ncbi.nlm.nih.gov/38875598/)

©Methun Kamruzzaman, Jack Heavey, Alexander Song, Matthew Bielskas, Parantapa Bhattacharya, Gregory Madden, Eili Klein, Xinwei Deng, Anil Vullikanti. Originally published in JMIR AI (<https://ai.jmir.org>), 16.05.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR AI, is properly cited. The complete bibliographic information, a link to the original publication on <https://www.ai.jmir.org/>, as well as this copyright and license information must be included.

Original Paper

Identifying Symptoms Prior to Pancreatic Ductal Adenocarcinoma Diagnosis in Real-World Care Settings: Natural Language Processing Approach

Fagen Xie¹, PhD; Jenny Chang¹, MPH; Tiffany Luong¹, MPH; Bechien Wu¹, MD, MPH; Eva Lustigova¹, MPH; Eva Shrader², MS; Wansu Chen¹, PhD

¹Department of Research and Evaluation, Kaiser Permanente Southern California, Pasadena, CA, United States

²Pancreatic Cancer Action Network, Manhattan Beach, CA, United States

Corresponding Author:

Fagen Xie, PhD

Department of Research and Evaluation

Kaiser Permanente Southern California

100 S Los Robles Avenue

Pasadena, CA, 91101

United States

Phone: 1 6265643294

Email: fagen.xie@kp.org

Abstract

Background: Pancreatic cancer is the third leading cause of cancer deaths in the United States. Pancreatic ductal adenocarcinoma (PDAC) is the most common form of pancreatic cancer, accounting for up to 90% of all cases. Patient-reported symptoms are often the triggers of cancer diagnosis and therefore, understanding the PDAC-associated symptoms and the timing of symptom onset could facilitate early detection of PDAC.

Objective: This paper aims to develop a natural language processing (NLP) algorithm to capture symptoms associated with PDAC from clinical notes within a large integrated health care system.

Methods: We used unstructured data within 2 years prior to PDAC diagnosis between 2010 and 2019 and among matched patients without PDAC to identify 17 PDAC-related symptoms. Related terms and phrases were first compiled from publicly available resources and then recursively reviewed and enriched with input from clinicians and chart review. A computerized NLP algorithm was iteratively developed and fine-trained via multiple rounds of chart review followed by adjudication. Finally, the developed algorithm was applied to the validation data set to assess performance and to the study implementation notes.

Results: A total of 408,147 and 709,789 notes were retrieved from 2611 patients with PDAC and 10,085 matched patients without PDAC, respectively. In descending order, the symptom distribution of the study implementation notes ranged from 4.98% for abdominal or epigastric pain to 0.05% for upper extremity deep vein thrombosis in the PDAC group, and from 1.75% for back pain to 0.01% for pale stool in the non-PDAC group. Validation of the NLP algorithm against adjudicated chart review results of 1000 notes showed that precision ranged from 98.9% (jaundice) to 84% (upper extremity deep vein thrombosis), recall ranged from 98.1% (weight loss) to 82.8% (epigastric bloating), and F_1 -scores ranged from 0.97 (jaundice) to 0.86 (depression).

Conclusions: The developed and validated NLP algorithm could be used for the early detection of PDAC.

(JMIR AI 2024;3:e51240) doi:[10.2196/51240](https://doi.org/10.2196/51240)

KEYWORDS

cancer; pancreatic ductal adenocarcinoma; symptom; clinical note; electronic health record; natural language processing; computerized algorithm; pancreatic cancer; cancer death; abdominal pain; pain; validation; detection; pancreas

Introduction

Pancreatic cancer is the third leading cause of cancer deaths in the United States, with 50,550 estimated deaths in 2023 [1].

Pancreatic ductal adenocarcinoma (PDAC), which accounts for 90% of pancreatic cancer cases, is the most common form of pancreatic cancer. The age- and sex-adjusted incidence has continued to increase, reaching 13.3 per 100,000 in 2015-2019,

and the overall 5-year survival remains poor at only 12.5% [2]. Despite technological advances, diagnosis of pancreatic cancer remains very late, with more than 50% of patients having distant metastases at the time of diagnosis [2-4].

Patient-reported symptoms are often the trigger for evaluation that eventually leads to a diagnosis of pancreatic cancer [5,6]. The reported prevalence of symptoms associated with PDAC has largely varied due to many factors, such as study design and data sources [6-10]. Additionally, previously published studies have been based on patient surveys [6,7] or structured electronic health records (EHRs) [8-10]. However, structured data can be inaccurate [11,12] and incomplete [13], especially for signs and symptoms. On the other hand, signs and symptoms are frequently collected and documented in the clinical notes by care providers via free text within the EHRs. Therefore, extracting signs and symptoms from clinical notes offers a key opportunity for the early detection of pancreatic cancer, which can lead to more timely interventions that improve survival.

Identification of PDAC-related symptoms from clinical notes based on EHRs is a challenge because signs or symptoms are typically not well-documented in a structured format within an EHR system, and specific techniques are required for data processing and analysis. Natural language processing (NLP), a field of computer-based methods aimed at standardizing and analyzing free text, processes unstructured data through information extraction from natural language and semantic representation learning for information retrieval, classifications, and predictions [14]. Numerous innovative NLP applications have been developed across various clinical domains in support of medical research, public health surveillance, clinical decision making, and outcome predictions [15-19]. Early NLP applications have largely focused on rule-based approaches [15,16], while recent NLP applications utilize state-of-the-art machine learning [17] or deep learning approaches via transformer learning models [18-20]. Rule-based NLP techniques have been widely used to extract signs and symptoms from free-text narratives in past years [21-26]. To the best of our knowledge, we are not aware of previous studies systematically analyzing pancreatic cancer-related symptoms from clinical notes via NLP. The purpose of this study is to develop and validate a comprehensive NLP algorithm and process to effectively identify PDAC-related symptoms prior to diagnosis within a large integrated health system.

Methods

Study Setting

Kaiser Permanente Southern California (KPSC) is an integrated health care system providing comprehensive medical services to over 4.8 million members across 15 large medical centers and more than 250 medical offices throughout the Southern California region. The demographic characteristics of KPSC members are diverse and largely representative of the residents in Southern California [27]. Members obtain their health insurance through group plans, individual plans, and Medicare and Medicaid programs and represent >260 ethnicities and >150 spoken languages. KPSC's extensive EHR data contains individual-level structured data (ie, diagnosis codes, procedure codes, medications, immunization records, laboratory results, and pregnancy episodes and outcomes) and unstructured data (ie, free-text clinical notes, radiology reports, pathology reports, imaging, and videos). KPSC's EHR covers all medical visits across all health care settings (eg, outpatient, inpatient, and emergency department). Clinical care of KPSC members provided by external contracted providers is captured in the EHR through reimbursement claim requests.

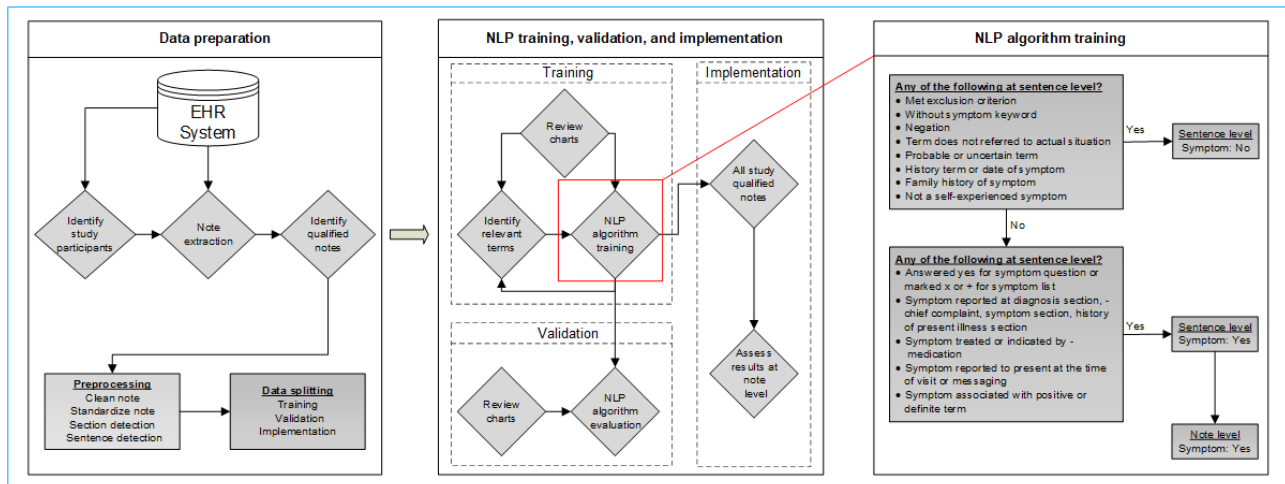
Ethical Considerations

The study protocol was reviewed and approved by the KPSC Institutional Review Board (approval no. 12849) with a waiver of the requirement for informed consent.

Study Population Identification

This study was a nested case-control study of KPSC patients aged 18-84 years between 2010 and 2019. Patients diagnosed with PDAC were identified through KPSC's cancer registry. Patients with a history of acute or chronic pancreatitis, without a clinic-based visit within 3 to 24 months prior to the diagnosis, with chemotherapy or infusion treatment, or with less than 20 months of health plan enrollment or pregnancy within 2 years prior to the diagnosis date were excluded. Among the patients with PDAC, the date of diagnosis was defined as the index date. For each PDAC case, up to 4 controls were selected from a group of patients without PDAC on the index date of the matched cases. Controls could develop PDAC 1 year after the index date. The above study criteria identified a total of 2611 eligible patients with PDAC and 10,085 corresponding matched patients without PDAC during the study period. The study participant identification and NLP process is shown in [Figure 1](#).

Figure 1. Schematic diagram of the NLP algorithm to identify the pancreatic ductal adenocarcinoma–related symptoms. EHR: electronic health record; NLP: natural language processing.



PDAC Symptom Selection

We initially identified 24 PDAC-related symptoms based on literature reviews and clinicians' input. A survey was conducted among the Consortium for the Study of Pancreatitis, Diabetes, and Pancreatic Cancer working group members [28] to determine the relative importance of the 24 potential symptoms. Based on the ranking of importance, a total of 17 symptoms were finally selected. In this study, we considered abdominal pain and epigastric pain as a combined symptom (abdominal or epigastric pain) and anorexia and early satiety as a combined symptom of (anorexia or early satiety) due to the difficulty of distinguishing them in clinical notes or patient-provider communications. The deep vein thrombosis (DVT) symptom was included in our study because DVT risk is high in patients with pancreatic cancer [29], and the symptom was further delineated into upper and lower DVT.

PDAC Symptom Keyword Selection

First, we compiled a list of phrases or terms relevant to the 17 symptoms based on previous literature [21-23] or symptom ontologies in the Unified Medical Language System [30]. The list was then reviewed and enriched by the experienced study gastroenterologist and enhanced by manual data annotation processing (refer to "Data Annotation" subsection for details). In addition, we used a word embedding model, Word2vec [31,32], to capture possible relevant phrases and terms, including misspelled terms, for each symptom. The compiled comprehensive phrases and terms for these 17 symptoms are summarized in Table S1 in Multimedia Appendix 1. The PDAC symptoms can be determined by a single phrase or term except for the DVT symptom. The DVT symptom was determined by 3 sets of terms, which included location (eg, leg or arm), feeling or appearance (eg, pain or swollen), and laterality (eg, left or right), rather than a single phrase or term.

Extraction and Preprocessing of Study Notes

Clinical notes and patient communication messages (telephone or email) within 2 years prior to the index date of PDAC cases and their matched controls (referred to as "notes" hereafter) were extracted from the KPSC EHR system. Notes associated with certain medical encounters (eg, surgery), note types (eg,

patient instructions or anesthesia), and department specialties (eg, health education) were excluded from the analysis because symptoms of interest were unlikely to be present in these notes (Table S2 in Multimedia Appendix 1). The extracted notes were then preprocessed through the following steps: (1) lowercase conversion, sentence splitting, and word tokenization [33]; (2) removal of nondigital or nonletter characters except for spaces, periods, commas, question marks, colons, and semicolons; (3) standardization of abbreviated words; and (4) correction of misspelled words based on the Word2vec model supplemented by an internal spelling correction file developed in previous studies [23,25].

Training, Validation, and Implementation Data Sets

Our study involved 2 phases of training and validation. The first phase used the notes of 100 randomly selected PDAC cases. The second phase used a subset of notes from both PDAC cases and controls. Details of the sample selection for training and validation are summarized in Table S3 in Multimedia Appendix 1. Notes that were not used for training or validation formed the study implementation data set.

Data Annotation

Notes from both the training and validation data sets were manually reviewed by trained research annotators to indicate the presence of the 17 symptoms based on the established terms and phrases (Table S1 in Multimedia Appendix 1) and inclusion and exclusion criteria (Table S4 in Multimedia Appendix 1). The note annotation process was based on a computer-assisted approach. First, notes from the training and validation data sets were exported into a spreadsheet and the prespecified terms (Table S1 in Multimedia Appendix 1) were highlighted. Second, for each note, the annotators reviewed the notes to label the presence of each of the 17 symptoms. Third, any ambiguous notes were fully discussed during weekly study team meetings until a consensus was reached. Cases that were difficult to determine were reported to the study gastroenterologist for adjudication.

A subset of the training data set in the first phase ($n=2795$ notes) was double-reviewed (ie, 2 annotators independently reviewed the same set of notes). The results from the 2 annotators were

compared and inconsistencies between them were discussed until a consensus was reached. If the annotators did not reach a consensus, the note was reviewed and adjudicated by the study gastroenterologist.

Finally, the adjudicated results were documented as the gold standard for training and validation of the NLP algorithm.

NLP Algorithm Development

Algorithm development involved 2 phases of training. For each phase, we used the annotated training data set to develop or refine a rule-based computerized algorithm via an iterative process to determine the presence of the 17 symptoms in each note. First, the notes were analyzed based on the phrase or terms and patterns that indicated the presence or absence of each symptom (Table S1 in [Multimedia Appendix 1](#)). The algorithm was then processed to search for patterns of inclusion or exclusion to determine the status of each symptom (Table S4 in [Multimedia Appendix 1](#)). A list of negated terms (eg, “ruled out” or “negative for”), uncertain or probable terms (eg, “presumably”), definite terms (eg, “positive for”), history terms (eg, “several years ago”), non-patient person terms (eg, referring to a family member), and general description terms (eg, “please return to ED if you have any of the following symptoms”) were compiled from the training data sets. The compiled terms were enriched via the repeated test-revise strategy against the chart review results within each training subset until the algorithm performance reached an acceptable threshold (ie, positive predictive value [PPV]=90%). The discordant cases between the algorithm and manually annotated results for each subset were further reviewed and adjudicated among the annotators and study team until a consensus was reached.

Specifically, each symptom for each note was first determined at the sentence level based on the following criteria:

1. A sentence defaulted as “no” if any exclusion criterion in Table S4 in [Multimedia Appendix 1](#) was met.
2. The symptom was considered absent if the sentence met any of the following situations:
 - The sentence did not contain any defined terms listed in Table S1 in [Multimedia Appendix 1](#).
 - The negated description was associated with defined terms listed in Table S1 in [Multimedia Appendix 1](#). Examples included “patient denied vomiting/nausea,” “ruled out jaundice,” and “no pruritus.”
 - The description of the symptom did not refer to an actual situation. For example, “return if you experience epigastric bloating” and “glipizide side effects including loss of appetite, nausea, vomiting, weight gain.”
 - A probable or uncertain description was associated with the symptom. For example, “patient with anxiety and likely depression” and “patient informed that there may be pruritis or pain.”
 - The symptoms were associated with a historical term or date relative to the clinical note date. For example, “patient had abdominal pain two years ago” and “patient had jaundice in 2007.”
 - The symptom description was related to family history, such as “family history: mother anxiety” and “patient family history: daughter with depression.”

- Someone other than the patient had a symptom. For example, “my husband is in a deep depression” and “daughter-in-law has been stressed, poor appetite and less sleep.”
 - The symptom was described as treated by medication during hospitalization.
 - The sentence only consisted of a symptom term, so a decision could not be reached on whether this instance was positive for the symptom.
3. A symptom was classified as “yes” for any of the following situations:
 - The sentence contained a symptom of interest and the symptom was marked as “yes,” “x,” or “+”. A symptom was classified as “yes” if the response to a symptom question was affirmative or if the symptom was marked on the symptom list.
 - The symptom was listed under the diagnosis section (except for DVT), chief complaint section, symptom section, and history of present illness section of the clinical note. For example, “chief complaint: abdominal pain,” “primary encounter diagnosis anxiety disorder,” and “jaundice 782.4.”
 - The symptom was described as treated or indicated by medication within nonhospitalization encounters.
 - The symptom was documented or reported to be present at the time of visit or messaging. For example, “pt complaint of 55 lb weight loss since March 2009” and “patient here for several weeks of abdominal pain.”
 - The sentence contained a definite term associated with a symptom of interest. Examples included “positive for fatigue and weight loss,” “patient reports anorexia,” and “patient presents with anxiety, depression, insomnia.”
 4. The sentence-level results were then combined to form note-level results.
 - Classification at the note level was defined as “yes” if at least 1 sentence in the note was marked “yes”. Otherwise, it was classified as “no”.

The diagnosis of DVT itself was not considered a DVT symptom. Additionally, the bodily location (ie, source) of pain was considered when determining the presence of any symptom (such as DVT, back pain, or abdominal or epigastric pain). For example, pain *radiating from* the upper or lower extremity was considered a DVT symptom, whereas pain *radiating to* the upper or lower extremity was not. Similarly, pain that *radiated to* the back region was not counted as back pain, and pain that *radiated to* the abdomen or epigastric region was not counted as abdominal or epigastric pain.

Performance Evaluation

The results of the NLP algorithm against the validation data set were compared to the adjudicated chart review results notes. For each symptom, the numbers of true positive (TP), false positive (FP), true negative (TN), and false negative (FN) cases were used to estimate the sensitivity or recall, specificity, PPV or precision, negative predictive value (NPV), and overall F_1 -score, a harmonic balance measurement of PPV and

sensitivity. Sensitivity was defined as the number of TPs divided by the total number of symptoms ascertained by the chart reviews (TP+FN). PPV was defined as the number of TPs divided by the total number of symptoms identified by the computerized algorithm (TP+FP). Specificity was defined as the number of TNs divided by the total number of notes without symptoms ascertained by the chart reviews (TN+FP). NPV was defined as the number of TNs divided by the total number of notes identified by the computerized algorithm without symptoms (TN+FN). The F_1 -score was calculated as $(2 \times PPV \times sensitivity) / (PPV + sensitivity)$.

Interrater Reliability Analysis Among 2 Annotators

The agreement and kappa coefficient against the double-annotated subset were calculated to assess the interrater reliability among the annotators.

Discrepancy Analysis

For each symptom, discordant results between the NLP algorithm and adjudicated chart review against the validation data set were analyzed. Both FP and FN scenarios were summarized in detail.

Implementation of the NLP Algorithm

The validated computerized algorithm was implemented via Python programming on a Linux server to process the qualified

study notes with the exception of training and validation notes. For each symptom, the process created the results of each note at the sentence level and note level for summary analysis.

Results

Statistics of the Study Notes

A total of 408,147 and 709,789 notes were retrieved for 2611 PDAC cases and 10,085 matched controls, respectively. The distribution of the notes and patient demographics are summarized in [Table 1](#). Compared to patients without PDAC, patients with PDAC were older and more likely to be men (PDAC cases: mean 69.2, SD 9.1 years of age and n=1328, 50.9% men; controls: mean 48.6, SD 17.2 years of age and n=4681, 46.4% men). A total of 3,827,166 sentences and 69,455,767 word tokens were derived from notes belonging to patients with PDAC. The corresponding numbers were 5,880,717 sentences and 102,358,031 word token for patients without PDAC. Both the average number of notes per patient and average words per note were higher for patients with PDAC (notes per patient: mean 156.3, SD 138.3; words per note: mean 170.2, SD 319.2) compared to patients without PDAC (notes per patient: mean 70.4, SD 94.1; words per note: mean 144.2, SD 263.6).

Table 1. Description of the study population and the associated data sets.

	PDAC ^a (n=2611)	Non-PDAC (n=10,085)
Age (years), mean (SD)	69.2 (9.1)	48.6 (17.2)
Gender: women, n (%)	1283 (49.1)	5404 (53.6)
Gender: men, n (%)	1328 (50.9)	4681 (46.4)
Total clinical notes, n	408,147	709,789
Total sentences, n	3,827,166	5,880,717
Total word tokens, n	69,455,767	102,358,031
Notes per patient, mean (SD)	156.3 (138.3)	70.4 (94.1)
Sentences per clinical note, mean (SD)	9.4 (15.7)	8.3 (13.9)
Words per clinical note, mean (SD)	170.2 (319.2)	144.2 (263.6)

^aPDAC: pancreatic ductal adenocarcinoma.

Interrater Reliability of 2 Annotators

The agreement and kappa coefficient between 2 annotators for a subset of notes (n=2795) is summarized in [Table S5](#) in [Multimedia Appendix 1](#). The agreement ranged from 98.82% (abdominal or epigastric pain) to 99.96% (upper extremity DVT), while the kappa coefficient ranged from 0.6 (insomnia) to 0.91 (abdominal or epigastric pain).

Validation of the NLP Algorithm

[Table 2](#) summarizes the performance of the computerized NLP algorithm against the adjudicated chart review results of 1000

notes based on the validation data set. In descending order, the precision (PPV) of the algorithms ranged from 98.9% (jaundice) to 84% (lower extremity DVT), recall (sensitivity) ranged from 98.1% (weight loss) to 82.8% (epigastric bloating), specificity ranged from 99.9% (epigastric bloating, jaundice, and pruritus) to 98.9% (depression), NPV ranged from 99.9% (lower extremity DVT) to 98.1% (abdominal or epigastric pain and back pain), and the F_1 -score ranged from 0.97 (jaundice) to 0.87 (depression).

Table 2. The computerized model's performance against the adjudicated chart review results in the validation data set (n=1000).

Symptoms	TP ^a (n)	TN ^b (n)	FP ^c (n)	FN ^d (n)	Sensitivity (%)	PPV ^e (%)	Specificity (%)	NPV ^f (%)	F ₁ -score
Gastrointestinal symptoms									
Abdominal or epigastric pain	156	824	4	16	90.7	97.5	99.5	98.1	0.94
Anorexia or early satiety	78	909	2	11	87.6	97.5	99.8	98.8	0.92
Dark urine	51	938	3	8	86.4	94.4	99.7	99.2	0.90
Epigastric bloating	53	935	1	11	82.8	98.2	99.9	98.8	0.90
Nausea or vomiting ^g	97	820	3	7	93.3	97	99.6	99.2	0.95
Pale stool	40	949	5	6	87	88.9	99.5	99.4	0.88
Systemic symptoms									
Back pain	95	882	6	17	84.8	94.1	99.3	98.1	0.89
Fatigue	105	883	2	10	91.3	98.1	99.8	98.9	0.95
Jaundice	90	905	1	4	95.7	98.9	99.9	99.6	0.97
Malaise	52	941	2	5	91.2	96.3	99.8	99.5	0.94
Pruritus	27	970	1	2	93.1	96.4	99.9	99.8	0.95
Weight loss	101	886	11	2	98.1	90.2	99.8	99.8	0.94
Mental symptoms									
Anxiety	79	911	3	7	91.9	96.3	99.7	99.2	0.94
Depression	83	892	10	15	84.7	89.3	98.9	98.3	0.87
Insomnia	62	925	7	6	91.2	89.9	99.3	99.4	0.91
Vascular conditions									
Lower extremity DVT ^h symptom	19	977	3	1	95	86.4	99.7	99.9	0.91
Upper extremity DVT symptom	21	972	4	3	87.5	84	99.6	99.7	0.86

^aTP: true positive.

^bTN: true negative.

^cFP: false positive.

^dFN: false negative.

^ePPV: positive predicted value.

^fNPV: Negative predicted value.

^gHospital encounter notes were excluded with the exception of emergency notes.

^hDVT: deep vein thrombosis.

Discrepancy Analysis

The discrepancy analysis is summarized in Table S6 in [Multimedia Appendix 1](#). The most common scenarios that resulted in FPs were failure of exclusion of the symptoms described in the patient medical problem list, failure of exclusion of symptoms from instructions, failure of negation, or failure of exclusion of a symptom from past medical history. The most common scenarios for FNs were false negation, missing specific terms or patterns of terms in the search list, false classification of past history symptoms, or false exclusion of symptoms described in relevant medication instructions.

Implementation of the NLP Algorithm

[Table 3](#) summarizes the symptoms identified by the validated NLP algorithms based on the implementation data set. Of the 393,003 and 708,489 notes belonging to PDAC and non-PDAC patients, respectively, at least 1 symptom was identified in 52,803 (13.44%) and 56,552 (7.98%) notes, respectively. The presence of symptoms ranged (in descending order) from 4.98% (abdominal or epigastric pain) to 0.05% (upper extremity DVT) in patients with PDAC and from 1.75% (back pain) to 0.01% (pale stool) in the patients without PDAC.

Table 3. Presence of symptoms identified by the computerized algorithms based on the implementation data set at the clinical note level.

Symptom	Clinical notes from patients with PDAC ^a , n (%) (n=393,003)	Clinical notes from patients without PDAC, n (%) (n=708,489)
Any of 17 symptoms	52,803 (13.44)	56,552 (7.98)
Gastrointestinal symptoms		
Abdominal or epigastric pain	19,582 (4.98)	11,274 (1.59)
Anorexia or early satiety	4393 (1.12)	1626 (0.23)
Dark urine	1511 (0.38)	121 (0.02)
Epigastric bloating	3217 (0.82)	1665 (0.24)
Nausea or vomiting	7754 (1.97)	7429 (1.05)
Pale stool	875 (0.22)	35 (0.01)
Systemic symptoms		
Back pain	8407 (2.14)	12,416 (1.75)
Fatigue	7170 (1.82)	9621 (1.36)
Jaundice	9118 (2.32)	305 (0.04)
Malaise	2984 (0.76)	4162 (0.59)
Pruritus	1872 (0.48)	622 (0.09)
Weight loss	8001 (2.04)	2619 (0.37)
Mental symptoms		
Anxiety	3924 (1)	10,843 (1.53)
Depression	4995 (1.27)	10,810 (1.53)
Insomnia	2228 (0.57)	4159 (0.59)
Vascular conditions		
Lower extremity DVT ^b symptom	807 (0.21)	1465 (0.21)
Upper extremity DVT symptom	215 (0.05)	719 (0.1)

^aPDAC: pancreatic ductal adenocarcinoma.

^bDVT: deep vein thrombosis.

Discussion

In this study, we developed computerized NLP algorithms to identify 17 symptoms that were documented prior to PDAC diagnosis from clinical notes and patient-provider communication emails. To our knowledge, this is the first study to systematically identify a set of symptoms related to PDAC using NLP. When assessed against the manually annotated results, the algorithm achieved a reasonable performance, with recall (sensitivity) ranging from 82.6% to 98.1% and precision (PPV) ranging from 84% to 98.9%.

Accurate extraction of symptoms embedded in free-text notes posed a significant challenge. First, the symptoms might be described in various portions of the notes. For example, symptoms might be embedded under past medical history, review of systems, the patient's medical problem list, instructions, sign and symptom warnings, questionnaires, checklists, lab orders and tests, medications, procedures, diagnosis, or chief complaints. Second, health care providers might copy and paste information from previous notes. In addition, we would like to highlight some specific challenges.

First, a negated term could sometimes apply to only 1 symptom or to multiple symptoms after negation (eg, no coughing, no chest pain, no abdomen pain; denies nausea or vomiting, diarrhea, constipation, abdominal pain). Second, the defined rules might not address all scenarios. For example, one of our defined rules for abdominal pain required the word "pain" and the body location to be within a 5-word distance. If the words for body location (eg, abdomen) and "pain" were separated by more than 5 words, the sentence was marked "no" for abdominal pain. Third, we found that some symptom terms could have different meanings, which caused FPs. For example, the phrase "lower bp" for back pain could also mean lower blood pressure, and the fatigue term "exhausted" could refer to either physical or mental exhaustion. Fourth, some exclusion criteria, as shown in Table S3 in [Multimedia Appendix 1](#) (eg, exclude localized itching for pruritus), also caused potential misclassification.

The data annotation process was tedious and time-consuming. The following lessons learned could benefit the medical research community. First, set up a training period for chart annotators and study investigators with medical backgrounds to review at least several hundred notes (the same notes for all the annotators). This step would not only allow the chart annotators

to be trained for the process but also would identify potential issues that might arise during the formal review process. Second, develop a chart annotation document that would include the detailed inclusion and exclusion criteria to be used for the annotation. The document should define specific types of notes (eg, mental health progress notes) or sections of the notes (eg, “past medical history” or “history of present illness”) to be reviewed or to be skipped. The document should also outline rules to determine the presence or absence of the conditions of interest. For example, if a patient experienced abdominal pain at home but did not experience pain at the time of the visit. Such rules are study-specific, but they need to be considered thoroughly and documented.

Advanced transformer language models, including bidirectional encoder representations from transformers (BERT) [20], clinical BERT [34], BioBERT [35], and BERT for EHRs (BEHRT) [36], have gained popularity in research involving NLP. These NLP language models offer the advantage of contextual understanding through embedding representations, allowing the developed algorithms to capture the meaning and intricate relationships within the text and enhance the accuracy of the analysis. They have been widely used for analyzing information from unstructured notes in the health care domain [18,19,37]. Research in this area in future work is warranted to further boost the performance of PDAC-related symptoms, especially for these lower performances via the rule-based approach.

Our study acknowledged several potential limitations. First, the completeness and accuracy of the extracted symptoms depended

on the information documented in the EHR system. Incomplete or inaccurate documentation of symptoms could lead to bias. Second, although our training process was quite comprehensive and included a relatively large number of notes, the rules and lexicons built based on the training data sets were still not highly comprehensive, as summarized in the discrepancy analysis. Therefore, a more extensive sample could be used to enhance the rules and lexicons if applied in other populations in the future, especially for rare symptoms. Third, a few terms or phrases could indicate meanings other than the symptom of interest (eg, “patient has exhausted all conservative measures” or “patient complaint of lower bp than usual”). Additional contexts with these terms would be required to determine the actual meaning. Fourth, for symptoms involving body location, such as abdominal pain and back pain, the allowed distance between the location and the symptom could sometimes lead to the misclassification of TP cases. Lastly, when applied to other health care systems and settings, the developed computerized algorithms might require modifications due to variations in the format and presentation of clinical notes in different health care settings.

In conclusion, the developed computerized algorithm and process could effectively identify relevant symptoms prior to PDAC diagnosis based on unstructured notes in a real-world care setting. This algorithm and process could be used to support the early detection of pancreatic cancer if implemented within a health care system to automatically identify patients with PDAC-related symptoms, especially those with PDAC-specific symptoms.

Acknowledgments

This study was supported by The Pancreatic Cancer Action Network. The opinions expressed are solely those of the authors and do not necessarily reflect the official views of the funding agency. The authors thank the survey participants from the Consortium for the Study of Pancreatitis, Diabetes, and Pancreatic Cancer working group to determine the PDAC-related symptoms. The authors thank the patients of Kaiser Permanente Southern California for helping to improve care through the use of information collected through our electronic health record systems.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Supplemental tables.

[DOCX File, 51 KB - ai_v3i1e51240_app1.docx]

References

1. American Cancer Society. URL: <https://www.cancer.org/content/dam/cancer-org/research/cancer-facts-and-statistics/annual-cancer-facts-and-figures/2023/2023-cancer-facts-and-figures.pdf> [accessed 2023-07-23]
2. Cancer stat facts: pancreatic cancer. Surveillance, Epidemiology, and End Results. URL: <https://seer.cancer.gov/statfacts/html/pancreas.html> [accessed 2023-07-24]
3. Stathis A, Moore MJ. Advanced pancreatic carcinoma: current treatment and future challenges. *Nat Rev Clin Oncol* 2010 Mar;7(3):163-172. [doi: [10.1038/nrclinonc.2009.236](https://doi.org/10.1038/nrclinonc.2009.236)] [Medline: [20101258](https://pubmed.ncbi.nlm.nih.gov/20101258/)]
4. Zhang L, Sanagapalli S, Stoita A. Challenges in diagnosis of pancreatic cancer. *World J Gastroenterol* 2018 May 21;24(19):2047-2060 [FREE Full text] [doi: [10.3748/wjg.v24.i19.2047](https://doi.org/10.3748/wjg.v24.i19.2047)] [Medline: [29785074](https://pubmed.ncbi.nlm.nih.gov/29785074/)]
5. Risch HA, Yu H, Lu L, Kidd MS. Detectable symptomatology preceding the diagnosis of pancreatic cancer and absolute risk of pancreatic cancer diagnosis. *Am J Epidemiol* 2015 Jul 01;182(1):26-34 [FREE Full text] [doi: [10.1093/aje/kwv026](https://doi.org/10.1093/aje/kwv026)] [Medline: [26049860](https://pubmed.ncbi.nlm.nih.gov/26049860/)]

6. Holly EA, Chaliha I, Bracci PM, Gautam M. Signs and symptoms of pancreatic cancer: a population-based case-control study in the San Francisco Bay area. *Clin Gastroenterol Hepatol* 2004 Jun;2(6):510-517. [doi: [10.1016/s1542-3565\(04\)00171-5](https://doi.org/10.1016/s1542-3565(04)00171-5)] [Medline: [15181621](https://pubmed.ncbi.nlm.nih.gov/15181621/)]
7. Walter FM, Mills K, Mendonça SC, Abel GA, Basu B, Carroll N, et al. Symptoms and patient factors associated with diagnostic intervals for pancreatic cancer (SYMPTOM pancreatic study): a prospective cohort study. *Lancet Gastroenterol Hepatol* 2016 Dec;1(4):298-306. [doi: [10.1016/s2468-1253\(16\)30079-6](https://doi.org/10.1016/s2468-1253(16)30079-6)]
8. Stapley S, Peters TJ, Neal RD, Rose PW, Walter FM, Hamilton W. The risk of pancreatic cancer in symptomatic patients in primary care: a large case-control study using electronic records. *Br J Cancer* 2012 Jun 05;106(12):1940-1944 [FREE Full text] [doi: [10.1038/bjc.2012.190](https://doi.org/10.1038/bjc.2012.190)] [Medline: [22617126](https://pubmed.ncbi.nlm.nih.gov/22617126/)]
9. Keane MG, Horsfall L, Rait G, Pereira SP. A case-control study comparing the incidence of early symptoms in pancreatic and biliary tract cancer. *BMJ Open* 2014 Nov 19;4(11):e005720 [FREE Full text] [doi: [10.1136/bmjopen-2014-005720](https://doi.org/10.1136/bmjopen-2014-005720)] [Medline: [25410605](https://pubmed.ncbi.nlm.nih.gov/25410605/)]
10. Watanabe I, Sasaki S, Konishi M, Nakagohri T, Inoue K, Oda T, et al. Onset symptoms and tumor locations as prognostic factors of pancreatic cancer. *Pancreas* 2004 Mar;28(2):160-165. [doi: [10.1097/00006676-200403000-00007](https://doi.org/10.1097/00006676-200403000-00007)] [Medline: [15028948](https://pubmed.ncbi.nlm.nih.gov/15028948/)]
11. Hersh W, Weiner M, Embi P, Logan J, Payne P, Bernstam E. Caveats for the use of operational electronic health record data in comparative effectiveness research. *Med Care* 2013;51:S30-S37. [doi: [10.1097/mlr.0b013e31829b1dbd](https://doi.org/10.1097/mlr.0b013e31829b1dbd)]
12. Diaz-Garelli J, Strowd R, Wells B, Ahmed T, Merrill R, Topaloglu U. Lost in translation: diagnosis records show more inaccuracies after biopsy in oncology care EHRs. *AMIA Jt Summits Transl Sci Proc* 2019;2019:325-334 [FREE Full text] [Medline: [31258985](https://pubmed.ncbi.nlm.nih.gov/31258985/)]
13. Zheng C, Yu W, Xie F, Chen W, Mercado C, Sy LS, et al. The use of natural language processing to identify Tdap-related local reactions at five health care systems in the Vaccine Safety Datalink. *Int J Med Inform* 2019 Jul;127:27-34 [FREE Full text] [doi: [10.1016/j.ijmedinf.2019.04.009](https://doi.org/10.1016/j.ijmedinf.2019.04.009)] [Medline: [31128829](https://pubmed.ncbi.nlm.nih.gov/31128829/)]
14. Friedman C, Alderson PO, Austin JH, Cimino JJ, Johnson SB. A general natural-language text processor for clinical radiology. *J Am Med Inform Assoc* 1994;1(2):161-174 [FREE Full text] [doi: [10.1136/jamia.1994.95236146](https://doi.org/10.1136/jamia.1994.95236146)] [Medline: [7719797](https://pubmed.ncbi.nlm.nih.gov/7719797/)]
15. Savova GK, Masanz JJ, Ogren PV, Zheng J, Sohn S, Kipper-Schuler KC, et al. Mayo clinical text analysis and knowledge extraction system (cTAKES): architecture, component evaluation and applications. *J Am Med Inform Assoc* 2010;17(5):507-513 [FREE Full text] [doi: [10.1136/jamia.2009.001560](https://doi.org/10.1136/jamia.2009.001560)] [Medline: [20819853](https://pubmed.ncbi.nlm.nih.gov/20819853/)]
16. Crowley RS, Castine M, Mitchell K, Chavan G, McSherry T, Feldman M. caTIES: a grid based system for coding and retrieval of surgical pathology reports and tissue specimens in support of translational research. *J Am Med Inform Assoc* 2010;17(3):253-264 [FREE Full text] [doi: [10.1136/jamia.2009.002295](https://doi.org/10.1136/jamia.2009.002295)] [Medline: [20442142](https://pubmed.ncbi.nlm.nih.gov/20442142/)]
17. Gonzalez-Hernandez G, Sarker A, O'Connor K, Savova G. Capturing the patient's perspective: a review of advances in natural language processing of health-related text. *Yearb Med Inform* 2017 Sep 11;26(01):214-227. [doi: [10.15265/iy-2017-029](https://doi.org/10.15265/iy-2017-029)]
18. Lu Z, Sim J, Wang JX, Forrest CB, Krull KR, Srivastava D, et al. Natural language processing and machine learning methods to characterize unstructured patient-reported outcomes: validation study. *J Med Internet Res* 2021 Nov 03;23(11):e26777 [FREE Full text] [doi: [10.2196/26777](https://doi.org/10.2196/26777)] [Medline: [34730546](https://pubmed.ncbi.nlm.nih.gov/34730546/)]
19. Arnaud É, Elbattah M, Gignon M, Dequen G. Learning embeddings from free-text triage notes using pretrained transformer models. In: *Proceedings of the 15th International Joint Conference on Biomedical Engineering Systems and Technologies*. 2022 Presented at: BIOSTEC 2022; February 9-11, 2022; Online p. 835-841. [doi: [10.5220/0011012800003123](https://doi.org/10.5220/0011012800003123)]
20. Devlin J, Chang M, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. In: *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 2018 Presented at: Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies; June 1-6, 2018; New Orleans, Louisiana p. 4171-4186. [doi: [10.18653/v1/n18-3](https://doi.org/10.18653/v1/n18-3)]
21. Koleck T, Dreisbach C, Bourne P, Bakken S. Natural language processing of symptoms documented in free-text narratives of electronic health records: a systematic review. *J Am Med Inform Assoc* 2019 Apr 01;26(4):364-379 [FREE Full text] [doi: [10.1093/jamia/ocy173](https://doi.org/10.1093/jamia/ocy173)] [Medline: [30726935](https://pubmed.ncbi.nlm.nih.gov/30726935/)]
22. Iqbal E, Mallah R, Rhodes D, Wu H, Romero A, Chang N, et al. ADEPt, a semantically-enriched pipeline for extracting adverse drug events from free-text electronic health records. *PLoS One* 2017;12(11):e0187121 [FREE Full text] [doi: [10.1371/journal.pone.0187121](https://doi.org/10.1371/journal.pone.0187121)] [Medline: [29121053](https://pubmed.ncbi.nlm.nih.gov/29121053/)]
23. Malden DE, Tartof SY, Ackerson BK, Hong V, Skarbinski J, Yau V, et al. Natural language processing for improved characterization of COVID-19 symptoms: observational study of 350,000 patients in a large integrated health care system. *JMIR Public Health Surveill* 2022 Dec 30;8(12):e41529 [FREE Full text] [doi: [10.2196/41529](https://doi.org/10.2196/41529)] [Medline: [36446133](https://pubmed.ncbi.nlm.nih.gov/36446133/)]
24. Matheny ME, Fitzhenry F, Speroff T, Green JK, Griffith ML, Vasilevskis EE, et al. Detection of infectious symptoms from VA emergency department and primary care clinical documentation. *Int J Med Inform* 2012 Mar;81(3):143-156. [doi: [10.1016/j.ijmedinf.2011.11.005](https://doi.org/10.1016/j.ijmedinf.2011.11.005)] [Medline: [22244191](https://pubmed.ncbi.nlm.nih.gov/22244191/)]
25. Zeiger RS, Xie F, Schatz M, Hong BD, Weaver JP, Bali V, et al. Prevalence and characteristics of chronic cough in adults identified by administrative data. *TPJ* 2020 Dec;24(5):1-14. [doi: [10.7812/tpp/20.022](https://doi.org/10.7812/tpp/20.022)]

26. Wang J, Abu-El-Rub N, Gray J, Pham H, Zhou Y, Manion F, et al. COVID-19 SignSym: a fast adaptation of a general clinical NLP tool to identify and normalize COVID-19 signs and symptoms to OMOP common data model. *J Am Med Inform Assoc* 2021 Jun 12;28(6):1275-1283 [FREE Full text] [doi: [10.1093/jamia/ocab015](https://doi.org/10.1093/jamia/ocab015)] [Medline: [33674830](https://pubmed.ncbi.nlm.nih.gov/33674830/)]
27. Koebnick C, Langer-Gould AM, Gould MK, Chao CR, Iyer R, Smith N, et al. Sociodemographic characteristics of members of a large, integrated health care system: comparison with US Census Bureau data. *Perm J* 2012 Sep 01;16(3):37-41. [doi: [10.7812/tpp/12-031](https://doi.org/10.7812/tpp/12-031)]
28. Steering committee of the PanCAN's EDI project. Pancreatic Cancer Action Network. URL: <https://pancan.org/research/early-detection-initiative/> [accessed 2023-05-03]
29. Johnson M, Sproule M, Paul J. The prevalence and associated variables of deep venous thrombosis in patients with advanced cancer. *Clin Oncol (R Coll Radiol)* 1999;11(2):105-110. [doi: [10.1053/clon.1999.9023](https://doi.org/10.1053/clon.1999.9023)] [Medline: [10378636](https://pubmed.ncbi.nlm.nih.gov/10378636/)]
30. Griffon N, Chebil W, Rollin L, Kerdelhue G, Thirion B, Gehanno J, et al. Performance evaluation of Unified Medical Language System's synonyms expansion to query PubMed. *BMC Med Inform Decis Mak* 2012 Feb 29;12:12 [FREE Full text] [doi: [10.1186/1472-6947-12-12](https://doi.org/10.1186/1472-6947-12-12)] [Medline: [22376010](https://pubmed.ncbi.nlm.nih.gov/22376010/)]
31. Goldberg Y, Levy O. word2vec explained: deriving Mikolov et al.'s negative-sampling word-embedding method. arXiv Preprint posted online on February 15, 2014. [FREE Full text]
32. Pennington J, Socher R, Manning C. Glove: global vectors for word representation. In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. 2014 Presented at: EMNLP 2014; October 25-29, 2014; Doha, Qatar. [doi: [10.3115/v1/d14-1162](https://doi.org/10.3115/v1/d14-1162)]
33. Loper E, Bird S. NLTK: the natural language toolkit. In: *Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics*. 2002 Presented at: TMTNLP 2002; July 7, 2002; Philadelphia, PA. [doi: [10.3115/1118108.1118117](https://doi.org/10.3115/1118108.1118117)]
34. Alsentzer E, Murphy J, Boag W, Weng W, Jindi D, Naumann T. Publicly available clinical BERT embeddings. In: *Proceedings of the 2nd Clinical Natural Language Processing Workshop*. 2019 Presented at: 2nd Clinical Natural Language Processing Workshop; June 7, 2019; Minneapolis, Minnesota. [doi: [10.18653/v1/w19-1909](https://doi.org/10.18653/v1/w19-1909)]
35. Lee J, Yoon W, Kim S, Kim D, Kim S, So C, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* 2020 Feb 15;36(4):1234-1240 [FREE Full text] [doi: [10.1093/bioinformatics/btz682](https://doi.org/10.1093/bioinformatics/btz682)] [Medline: [31501885](https://pubmed.ncbi.nlm.nih.gov/31501885/)]
36. Li Y, Rao S, Solares J, Hassaine A, Ramakrishnan R, Canoy D, et al. BEHRT: transformer for electronic health records. *Sci Rep* 2020 Apr 28;10(1):7155 [FREE Full text] [doi: [10.1038/s41598-020-62922-y](https://doi.org/10.1038/s41598-020-62922-y)] [Medline: [32346050](https://pubmed.ncbi.nlm.nih.gov/32346050/)]
37. Hossain E, Rana R, Higgins N, Soar J, Barua PD, Pisani AR, et al. Natural language processing in electronic health records in relation to healthcare decision-making: a systematic review. *Comput Biol Med* 2023 Mar;155:106649. [doi: [10.1016/j.compbiomed.2023.106649](https://doi.org/10.1016/j.compbiomed.2023.106649)] [Medline: [36805219](https://pubmed.ncbi.nlm.nih.gov/36805219/)]

Abbreviations

BERT: bidirectional encoder representations from transformers

DVT: deep vein thrombosis

EHR: electronic health record

FN: false negative

FP: false positive

KPSC: Kaiser Permanente Southern California

NLP: natural language processing

NPV: negative predictive value

PDAC: pancreatic ductal adenocarcinoma

PPV: positive predictive value

TN: true negative

TP: true positive

Edited by K El Emam, B Malin; submitted 26.07.23; peer-reviewed by B Sens, M Elbattah, Y Khan; comments to author 17.11.23; revised version received 08.12.23; accepted 16.12.23; published 15.01.24.

Please cite as:

Xie F, Chang J, Luong T, Wu B, Lustigova E, Shrader E, Chen W

Identifying Symptoms Prior to Pancreatic Ductal Adenocarcinoma Diagnosis in Real-World Care Settings: Natural Language Processing Approach

JMIR AI 2024;3:e51240

URL: <https://ai.jmir.org/2024/1/e51240>

doi: [10.2196/51240](https://doi.org/10.2196/51240)

PMID: [38875566](https://pubmed.ncbi.nlm.nih.gov/38875566/)

©Fagen Xie, Jenny Chang, Tiffany Luong, Bechien Wu, Eva Lustigova, Eva Shrader, Wansu Chen. Originally published in JMIR AI (<https://ai.jmir.org>), 15.01.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR AI, is properly cited. The complete bibliographic information, a link to the original publication on <https://www.ai.jmir.org/>, as well as this copyright and license information must be included.

Original Paper

Sepsis Prediction at Emergency Department Triage Using Natural Language Processing: Retrospective Cohort Study

Felix Brann¹, BSc; Nicholas William Sterling¹, MD, PhD, MS; Stephanie O Frisch¹, PhD, RN; Justin D Schrager^{1,2}, MD, MPH

¹Vital Software, Inc, Claymont, DE, United States

²Department of Emergency Medicine, Emory University School of Medicine, Atlanta, GA, United States

Corresponding Author:

Justin D Schrager, MD, MPH

Department of Emergency Medicine

Emory University School of Medicine

531 Asbury Circle

Annex Building N340

Atlanta, GA, 30322

United States

Phone: 1 404 778 5975

Email: justin@vitaler.com

Abstract

Background: Despite its high lethality, sepsis can be difficult to detect on initial presentation to the emergency department (ED). Machine learning–based tools may provide avenues for earlier detection and lifesaving intervention.

Objective: The study aimed to predict sepsis at the time of ED triage using natural language processing of nursing triage notes and available clinical data.

Methods: We constructed a retrospective cohort of all 1,234,434 consecutive ED encounters in 2015–2021 from 4 separate clinically heterogeneous academically affiliated EDs. After exclusion criteria were applied, the final cohort included 1,059,386 adult ED encounters. The primary outcome criteria for sepsis were presumed severe infection and acute organ dysfunction. After vectorization and dimensional reduction of triage notes and clinical data available at triage, a decision tree–based ensemble (time-of-triage) model was trained to predict sepsis using the training subset (n=950,921). A separate (comprehensive) model was trained using these data and laboratory data, as it became available at 1-hour intervals, after triage. Model performances were evaluated using the test (n=108,465) subset.

Results: Sepsis occurred in 35,318 encounters (incidence 3.45%). For sepsis prediction at the time of patient triage, using the primary definition, the area under the receiver operating characteristic curve (AUC) and macro F_1 -score for sepsis were 0.94 and 0.61, respectively. Sensitivity, specificity, and false positive rate were 0.87, 0.85, and 0.15, respectively. The time-of-triage model accurately predicted sepsis in 76% (1635/2150) of sepsis cases where sepsis screening was not initiated at triage and 97.5% (1630/1671) of cases where sepsis screening was initiated at triage. Positive and negative predictive values were 0.18 and 0.99, respectively. For sepsis prediction using laboratory data available each hour after ED arrival, the AUC peaked to 0.97 at 12 hours. Similar results were obtained when stratifying by hospital and when Centers for Disease Control and Prevention hospital toolkit for adult sepsis surveillance criteria were used to define sepsis. Among septic cases, sepsis was predicted in 36.1% (1375/3814), 49.9% (1902/3814), and 68.3% (2604/3814) of encounters, respectively, at 3, 2, and 1 hours prior to the first intravenous antibiotic order or where antibiotics were not ordered within the first 12 hours.

Conclusions: Sepsis can accurately be predicted at ED presentation using nursing triage notes and clinical information available at the time of triage. This indicates that machine learning can facilitate timely and reliable alerting for intervention. Free-text data can improve the performance of predictive modeling at the time of triage and throughout the ED course.

(JMIR AI 2024;3:e49784) doi:[10.2196/49784](https://doi.org/10.2196/49784)

KEYWORDS

natural language processing; machine learning; sepsis; emergency department; triage

Introduction

Background

Sepsis is a life-threatening condition caused by severe infection and dysregulated host response leading to acute organ dysfunction [1]. Affecting 32 million people and contributing to over 5 million deaths per year globally [2], sepsis is a leading cause of death in hospitalizations in the United States and worldwide [3,4]. Early antibiotics have been shown to improve survival [5], while each hour of delayed antibiotic administration has been associated with progressively increased mortality (7.6% increase per hour in septic shock) [6]. Patients who survive sepsis often have long-lasting health and social sequelae [7], and sepsis is ranked among the top 3 most costly conditions to treat in the hospital setting [8]. Accordingly, substantial efforts have been made to identify sepsis early in the hospital course [9]. To date, however, widely used clinical decision support tools that use rule-based methods for detecting sepsis have been limited by low sensitivity and specificity [10,11]. Such tools have been unable to earn clinician trust due to limited accuracy, false positives, and delayed alerts [12]. False positive alerts increase the cognitive load of providers and could expose patients to unnecessary antimicrobials. Moreover, current widely used electronic health record–based sepsis prediction tools have limited performance and often require several hours to elapse to achieve reasonable predictive use [12]. For example, a recent inpatient and intensive care unit (ICU)–based investigation of a commonly used sepsis alerting system showed that although existing systems can generate reasonably accurate sepsis alerts, the median time to notification was 7 hours and, even at that point, accuracy was limited [13]. Taken together, existing clinical decision support systems aimed at detecting sepsis do not provide sufficient accuracy or timeliness of sepsis prediction, resulting in lower adoption due to a lack of clinician trust.

Machine Learning in Sepsis Prediction

Artificial intelligence (AI)–based tools may hold promise to increase the accuracy and timeliness of sepsis prediction, which may allow for earlier delivery of critical interventions such as lifesaving antibiotics. Many of the most promising sepsis predictive algorithms have been limited to use in ICU settings [14], where patients have rich laboratory and imaging data sets and frequent physiologic monitoring. In contrast, accurate prediction of sepsis at initial emergency department (ED) presentation has remained elusive. Until recently, there was a paucity of technology that could make use of the full set of available data, particularly free-text triage notes, at the time of initial ED presentation. A recent study showed that sepsis prediction at the time of triage can be significantly improved using natural language processing (NLP) of free-text data [15].

ED Triage Assessment

When a patient presents to the ED, an initial triage assessment is usually performed by a triage nurse. The triage assessment includes a brief interview of the patient or those accompanying the patient to obtain a reason for presenting to the hospital ED. The content of this interview typically includes a very brief recounting of the patient's past medical history, relevant medications, family history, and social risk factors. The triage

nurse will typically also obtain vital signs (blood pressure, heart rate, temperature, respiratory rate, and oxygen saturation) and pain score. Finally, the triage nurse will assign a patient a triage acuity score. This process usually takes less than 10 minutes. The summation of this encounter is documented in real time, directly after the triage assessment, into the electronic medical record and includes a listing of the vital signs, triage acuity score, and a free-text nursing triage note.

The triage note is recorded into the electronic medical record, typically comprising 1-3 sentences regarding why the patient has presented to the ED and the nurse's summative impression of this initial assessment. This note is used as a starting point for downstream assessments by providers in the ED. The information contained in the triage note is useful, as it often contains rich data that are difficult to quantify in tabular form. This information is widely used and valued by the clinical staff. However, in its unstructured format, it is not typically used in clinical decision support algorithms and is often unused for several hours until the full provider assessment. We hypothesized that nursing triage notes, combined with other data available at initial ED presentation, could be used to accurately predict sepsis at the time of triage.

Goals of This Investigation

It was previously demonstrated that NLP of nursing triage notes at ED presentation could be used to predict hospital admission and ED resource use [16-18]. In this study, we aimed to demonstrate that an NLP-based model could be used to predict sepsis in adult patients based on the (1) health system sepsis committee and (2) Centers for Disease Control and Prevention (CDC) hospital toolkit for adult sepsis surveillance criteria [1].

Methods

Ethical Considerations

The research study protocol and procedures were reviewed and approved by the institutional review board (STUDY00000099).

Study Design and Setting

A retrospective cohort was constructed using electronic health record data from all 1,234,434 consecutive ED encounters (487,296 unique patients) in 2015-2020 from 4 separate clinically heterogeneous academically affiliated EDs. Hospital A is a community hospital in an urban setting having a patient volume of approximately 65,000 ED visits per year. Hospital B is a community hospital in a suburban setting having a volume of approximately 26,000 visits per year. Hospital C is a quaternary care academic medical setting in a major metropolitan area having an ED patient volume of approximately 48,000 visits per year. Hospital D is a community hospital in a suburban setting having a volume of approximately 36,000 visits per year.

Selection of Participants

Prior studies have suggested that overwhelming viral septicemia during the COVID-19 pandemic led to markedly increased false positive rates of sepsis screening tools [15]. These cases accounted for a substantial portion of ED visits during the initial months of 2020 [19] and led to a sharp decline in ED patient

volume [20]. Accordingly, we excluded encounters (n=94,739) from February 1, 2020, to August 1, 2020, and patients who had a diagnostic code of COVID-19 or positive COVID-19 laboratory test. Patients of 18 years and younger of age were excluded from the study (n=27,238), as defining sepsis in these patients is controversial, and they are often lost to follow-up after they are transferred for admission to pediatric hospitals. Patients whose date of birth or age was not available were also excluded (n=23,434) to ensure that the remaining cohort comprised only adult patients. We subsequently excluded encounters with missing triage notes (n=29,637). The final cohort of interest included 1,059,386 unique clinical encounters.

Sepsis Definition

The primary outcome of sepsis was defined as presumed severe infection and acute organ dysfunction, based on criteria described by the health system sepsis committee. To evaluate model performance against verified sepsis cases, the health system sepsis committee provided physician-reviewed sepsis labels for 7663 patients between June 1, 2019, and October 1, 2019. These cases were oversampled into the test data set. This definition of sepsis was projected onto the remaining data using clinical outcome variables. For sensitivity analyses of model performance, a secondary definition of sepsis was used, based on the US Centers for Medicare & Medicaid Services toolkit criteria [1]. Encounters were counted as sepsis, if they met criteria at any time during the ED course or hospital stay.

Natural Language Processing

NLP techniques have been developed to extract meaning from unstructured free-text data. One such technique is document vectorization. Documents can be transformed into numerical vectors that represent the key information they contain, allowing them to be used by numerical machine learning (ML) techniques.

To generate document embeddings for the nursing triage notes, a distilled BERT (Bidirectional Encoder Representation From Transformers) model pretrained using an unsupervised masked language modeling objective was used as a base. Unlike models pretrained using a causal language modeling objective such as Generative Pre-Trained Transformer, which only consider preceding tokens, BERT considers tokens to the right and left of the masked word [21].

The use of large models such as BERT is constrained by the computational resources required for training and inference. DistilBERT [22] is a lighter and faster language model that offers fewer constraints on computational resources, having a depth of only 6 layers, rather than 12, and with token-type embeddings and pooler removed. DistilBERT is trained to replicate the behavior of BERT using “teacher-student” learning, where BERT is the “teacher” and DistilBERT is the “student.” This allows for knowledge distillation in the pretraining phase while retaining 97% of language understanding and being 60% faster.

The base DistilBERT model was fine-tuned using the free textual data from nursing triage notes with the objective of predicting sepsis. We evaluated several pretrained document vectorization models, selecting the optimal one by calculating

fine-tuning performance on the training set. Nursing triage notes concatenated with Boolean clinical variables available at the time of triage (ie, high or low vital signs) were then passed through the fine-tuned DistilBERT model to produce document vectors representing the key information they contain. For the document vectors, we selected thresholds for the numeric values based on clinical knowledge and appended text based on the numeric values and those thresholds. Additionally, we developed manual mappings for known clinical abbreviations and converted them into the text. For example, “n/v/d” became “nausea, vomiting, and diarrhea.” The document vectors were then passed through a principal component analysis step to dimensionally reduce them from a length of 768 to 20 components.

Model Training and Testing

For the time-of-triage model, the triage note vectors were combined with other clinical data, such as age, sex, and maximum and minimum vital signs. For the prediction of sepsis after laboratory data availability, a separate comprehensive model was constructed that included the aforementioned variables and additional laboratory data.

While many sepsis indicators have clear unidirectional associations with sepsis risk (ie, heart rate, hypotension, and lactic acid), others can be bidirectional (ie, high or low temperature or white blood cell [WBC] count). In addition, triage note vectors may potentially have complex relationships with sepsis. Accordingly, a decision tree-based technique was chosen for model training over more traditional techniques, such as logistic regression. The combined vectors from the training data set were used to train a decision tree-based ensemble learning model (XGBoost [Extreme Gradient Boosting]) [23] to predict the likelihood of sepsis. The XGBoost model was trained to predict sepsis using the training subset (n=950,921). Model performance was evaluated using the test (n=108,465) subset.

Optimal hyperparameters for the time-of-triage model were determined via grid search. The time-of-triage model was trained using a maximum tree depth of 6, minimum child weight of 15, minimum split loss of 15, learning rate of 0.05, subsample ratio of 0.6, L1 regularization of 0, and L2 regularization of 1. After Bayesian hyperparameter optimization, the comprehensive model was trained using a maximum tree depth of 6, minimum child weight of 13, minimum split loss of 18, learning rate of 0.015, subsample ratio of 0.63, L1 regularization of 0.27, and L2 regularization of 1.87. We accounted for class imbalance by scaling the positive weight parameter to the inverse of the class distribution. Epoch-level evaluation was used to measure model performance during training and identify failing training runs. Heat maps to indicate word and subword importance were generated using the integrated gradients method on the constructed model inputs [24]. Word importance here was calculated on words and subwords returned by the tokenization method.

For analysis of sensitivity, specificity, and false positive rate of the time-of-triage model, a target threshold of model prediction score was selected based on optimizing for a maximal false positive rate of 0.15. For the comprehensive model, we derived

a classification threshold empirically, based on probability scores, and subsequently applied the threshold to target a maximum false positive rate of 0.1 at 12 hours after ED arrival. The thresholds were selected using model output scores from the training set and were applied to the test data set to evaluate clinical predictive performance metrics. The comprehensive model included known laboratory indicators of sepsis and end organ dysfunction, such as maximum and minimum WBC count, maximum lactic acid, minimum platelets, and maximum bilirubin and creatinine. Comprehensive model performance was evaluated using the test data set at every hour after ED arrival. Model performance was also evaluated at each hospital.

Sepsis Prediction Prior to the First Intravenous Antibiotic Order

To estimate how an AI sepsis prediction tool might impact the ordering of antibiotics, we computed the percentage of sepsis encounters that triggered a positive prediction of sepsis prior to antibiotics being ordered or not having antibiotics ordered within the first 12 hours of the encounter. To perform this analysis, we used encounters from the test data set. A dual-model approach was used to emulate sepsis alerting at the time of triage and then subsequently during the ED encounter. Sepsis prediction time was defined as the earlier of either the time-of-triage model or comprehensive model generating a positive prediction of sepsis.

Evaluation of Model Performances Among Clinically Undetected Sepsis Cases

To determine how the time-of-triage and comprehensive models may prevent missed sepsis, encounters with sepsis in the test data set were stratified by model prediction of sepsis- versus chart-based indicators of clinical sepsis suspicion. Predictive performance of the model was evaluated among patients who were septic and were or were not screened for sepsis at triage and defined as having either of the following order in less than 30 minutes after time of triage: (1) nursing-driven sepsis screening order set or (2) blood culture.

Results

Characteristics of the Study Patients

The total data set after exclusions consisted of 1,059,386 unique encounters from 487,296 patients. Sepsis occurred in 35,318 encounters (incidence 3.45%). Median time from arrival to first WBC count collection was 44.9 (IQR 26.2-79.3), 42.8 (IQR 25.6-73.3), and 44.8 (IQR 26.2-79.0) minutes across nonsepsis, sepsis, and all encounters, respectively. Demographic characteristics of the patients are available in [Table 1](#). Gender, race, and temperature were missing in 5.6% (57,082/1,059,386), 13.2% (87,284/1,059,386), and 0.2% (2034/1,059,386) of encounters, respectively. Respiratory rate, heart rate, oxygen saturation, and blood pressure were missing in 0.1% of encounters. Selected examples of triage notes of encounters where patients were septic are included in [Table S1](#) in [Multimedia Appendix 1](#).

Table 1. Demographic and clinical characteristics of patients across encounters.

	Total	Hospital A	Hospital B	Hospital C	Hospital D
Sepsis^a, n (%)	1,059,386 (100)	386,961 (36.5)	158,757 (15)	284,794 (26.9)	228,874 (21.6)
Primary	35,318 (3.3)	9533 (2.5)	3978 (2.5)	12,775 (4.5)	9032 (3.9)
Secondary	31,542 (3)	9128 (2.4)	3541 (2.2)	12,688 (4.5)	6185 (2.7)
Age (years), mean (SD)					
18-24	80,384 (7.6)	35,421 (9.2)	11,466 (7.2)	23,309 (8.2)	10,188 (4.5)
25-44	344,034 (32.5)	147,085 (38.0)	47,283 (29.8)	91,106 (32.0)	58,560 (25.6)
45-64	327,584 (30.9)	123,225 (31.8)	53,226 (33.5)	87,113 (30.6)	64,020 (28.0)
65-74	141,943 (13.4)	44,840 (11.6)	19,709 (12.4)	41,425 (14.5)	35,969 (15.7)
≥75	165,441 (15.6)	36,390 (9.4)	27,073 (17.1)	41,841 (14.7)	60,137 (26.3)
Sex, n (%)					
Female	579,798 (57.8)	208,230 (56.8)	90,599 (60.4)	160,710 (59.6)	120,259 (55.6)
Male	422,506 (42.2)	158,321 (43.1)	59,447 (39.6)	108,611 (40.3)	96,127 (44.4)
Race, n (%)					
Black	552,432 (50.6)	301,619 (75.6)	35,366 (21.7)	150,454 (51.3)	64,993 (27.6)
White	380,084 (34.8)	53,427 (13.3)	92,713 (56.8)	104,290 (35.6)	129,654 (56.6)
Other	39,586 (36.3)	5205 (1.3)	15,827 (9.7)	10,125 (3.5)	8429 (3.6)
Unreported	87,284 (8.2)	26,710 (6.9)	14,851 (9.4)	19,925 (7.0)	25,798 (11.3)
Vital signs					
Temperature (°C), mean (SD)	36.8 (0.5)	36.8 (0.5)	36.8 (0.5)	36.7 (0.6)	36.8 (0.5)
Heart rate (beats per minute), mean (SD)	85.6 (18.8)	86.2 (18.1)	84.5 (18.7)	85.9 (19.1)	84.8 (19.7)
Systolic BP ^b (mm Hg), mean (SD)	138.6 (26.7)	138.6 (26.9)	137.6 (24.4)	139.7 (28.9)	137.9 (24.9)
Diastolic BP (mm Hg), mean (SD)	80.0 (15.5)	80.8 (14.9)	80.2 (14.8)	80.5 (16.0)	77.8 (16.1)
SpO ₂ ^c (%), median (IQR)	98.0 (97-100)	98.0 (97-100)	98.0 (97-100)	98.0 (97-100)	99.0 (97-100)
Respiratory rate (breaths per minute), mean (SD)	18.0 (6.3)	18.2 (6.4)	18.0 (5.9)	18.1 (6.7)	17.8 (5.9)
Time to first WBC ^d count (minutes), median (IQR)	44.8 (26.5-80.3)	51.2 (27.3-85.0)	40.9 (20.8-62.8)	47.4 (32.4-90.3)	34.6 (23.1-73.0)

^aSepsis primary and secondary definitions based on the health system sepsis committee and Centers for Disease Control and Prevention hospital toolkit for adult sepsis surveillance criteria, respectively.

^bBP: blood pressure.

^cSpO₂: oxygen saturation.

^dWBC: white blood cell.

Time-of-Triage and Comprehensive Model Performances

Using the test data set, the time-of-triage model using information available at initial triage for sepsis prediction (primary criteria) demonstrated an area under the receiver operating characteristic curve (AUC) and macro F_1 -score of

0.94 and 0.61, respectively (Figure 1). Sensitivity, specificity, and false positive rate were 0.87, 0.85, and 0.15, respectively. Positive and negative predictive values were 0.18 and 0.99, respectively. Sample model output is available in Figure 2, depicted as heat maps applied to words and subwords of ED nursing triage notes to indicate positive, neutral, or negative contributions to sepsis prediction.

Figure 1. Receiver operating characteristic curve of sepsis prediction at the time of initial emergency department triage using free-text triage nursing notes and clinical data available at the time of triage. AUC: area under the receiver operating characteristic curve.

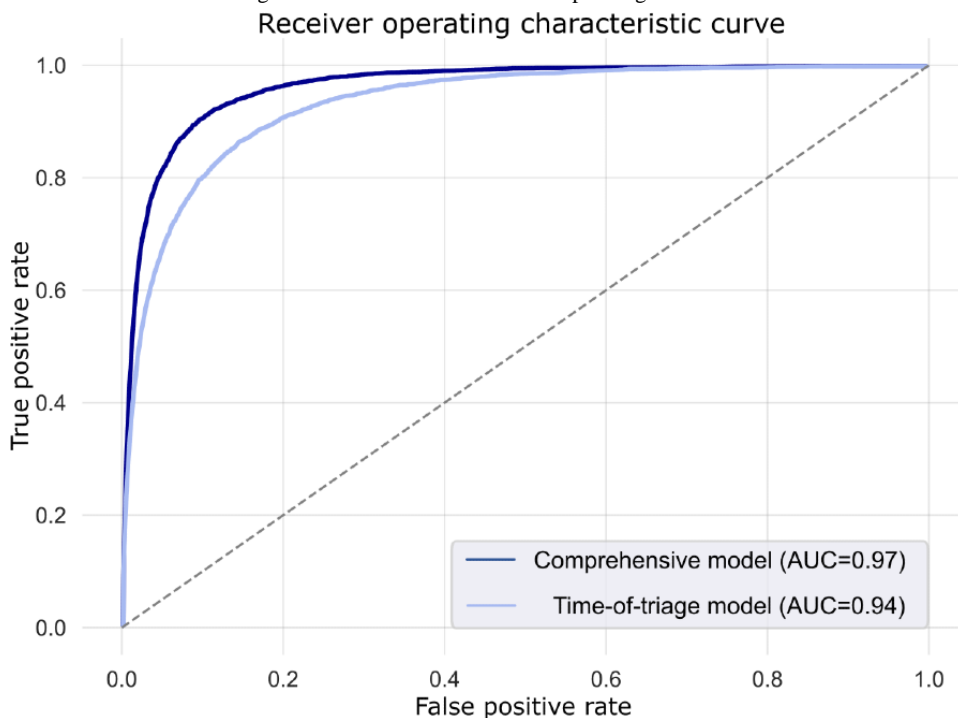


Figure 2. Heat maps applied to words and subwords of a sample of emergency department nursing triage notes to indicate relative contributions to sepsis prediction.



Incorporating data available after initial ED workup, the comprehensive model predicted sepsis based on primary criteria with an initial AUC, sensitivity, and specificity of 0.94, 0.72, and 0.94 at 1 hour after ED arrival, respectively; increasing to an AUC, sensitivity, and specificity of 0.96, 0.87, and 0.91 after 5 hours, respectively; and increasing to AUC, sensitivity, and specificity of 0.97, 0.91, and 0.90 at 12 hours after arrival,

respectively (Figure 3). Sensitivity, specificity, and false positive rate at 12 hours were 0.92, 0.89, and 0.11, respectively. Positive and negative predictive values at 12 hours were 0.25 and 0.99, respectively. Similar sepsis prediction results were obtained using the CDC hospital toolkit for adult sepsis surveillance criteria (Table 2) and when stratifying by hospital (Table S2 in Multimedia Appendix 1).

Figure 3. Sepsis predictive performance of the comprehensive model using a test data set, expressed as AUC, at each hour after emergency department arrival. AUC: area under the receiver operating characteristic curve.

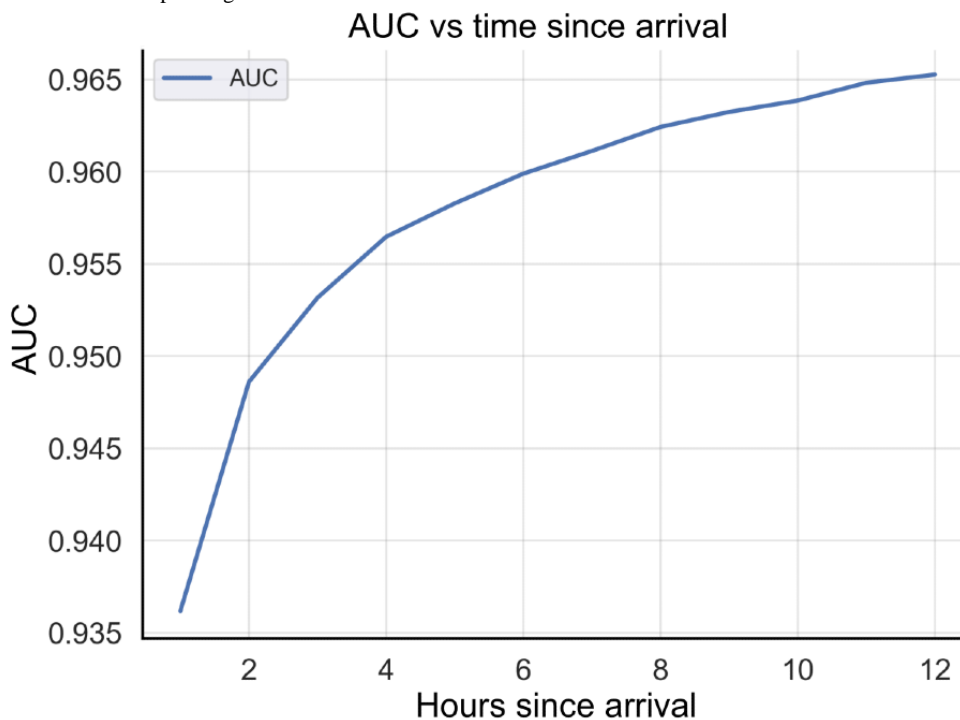


Table 2. Machine learning prediction of sepsis using data available at the time of emergency department (ED) triage (“time-of-triage” model) and all data available after ED workup (“comprehensive” model).

	Time-of-triage model	Comprehensive model
Primary sepsis criteria		
AUC ^a	0.94	0.97
Macro F_1	0.61	0.67
Sensitivity	0.87	0.91
Specificity	0.85	0.90
False positive rate	0.15	0.10
CDC^b hospital toolkit for adult sepsis surveillance		
AUC	0.92	0.96
Macro F_1	0.57	0.64
Sensitivity	0.86	0.91
Specificity	0.83	0.89
False positive rate	0.17	0.11

^aAUC: area under the receiver operating characteristic curve.

^bCDC: Centers for Disease Control and Prevention.

Model Performances Among Clinically Undetected Sepsis Cases

Sepsis screening initiated at triage was defined as having chart-based indicators of sepsis screening ordered within 30 minutes of triage (see Methods section). Within the test data set, there were 3821 encounters having sepsis. Among these, 1671 (43.7%) encounters had sepsis screening initiated at triage. The time-of-triage model accurately predicted sepsis in 76% (1635/2150) of sepsis cases where sepsis screening was not

initiated at triage and 97.5% (1630/1671) of cases where sepsis screening was initiated at triage.

Model Performances Among Critical Sepsis Cases

Among patients in the test data set who had sepsis and were ultimately placed on vasopressors or were admitted to the ICU, the time-of-triage model predicted sepsis in 97.9% (329/336) and 91.6% (832/908) encounters, respectively. The comprehensive model predicted sepsis in 100% (336/336) and 95.7% (869/908) encounters, respectively.

Sepsis Prediction Prior to the First Intravenous Antibiotic Order

We retrospectively evaluated the time of sepsis prediction in relation to the first intravenous antibiotic order using a dual-model approach (“time-of-triage” followed by “comprehensive” models). Among septic cases, sepsis was predicted in 36.1% (1375/3814), 49.9% (1902/3814), and 68.3% (2604/3814) of encounters at 3 hours, 2 hours, and 1 hour, respectively, prior to the first intravenous antibiotic order or where antibiotics were not ordered within the first 12 hours.

Model Performance Using Only the First Encounter per Patient

To ensure that model performance was not confounded by past encounters, we performed a sensitivity analysis using only the first encounter per patient in the test data set ($n=88,309$), excluding subsequent encounters. The time-of-triage model predicted sepsis with an AUC, sensitivity, specificity, and false positive rate of 0.94, 0.85, 0.86, and 0.14, respectively. The comprehensive model predicted sepsis at 12 hours with an AUC, sensitivity, specificity, and false positive rate of 0.97, 0.92, 0.90, and 0.10, respectively.

Analysis of Model Feature Importance

The importance of model features was analyzed by ranking the XGBoost feature importance scores from highest to lowest (Figure S1 in [Multimedia Appendix 1](#)). For both the time-of-triage (Figure S2 in [Multimedia Appendix 1](#)) and comprehensive (Figure S3 in [Multimedia Appendix 1](#)) models, the top features included elements of vital signs (ie, heart rate, temperature, blood pressure, and oxygen saturation) and triage note vectors. For the comprehensive model, the most important features additionally included laboratory metrics such as WBC count, creatinine, and lactic acid.

Discussion

Principal Findings

In this study, data from over 1 million patient encounters across 4 large metropolitan EDs were used to train an NLP-based ML model to detect sepsis at the time of patient presentation to the ED. We demonstrated that free-text nursing triage notes, combined with clinical variables at the time of triage, could be used to accurately predict the occurrence of sepsis at initial ED nursing triage. Moreover, we demonstrated that sepsis could be detected in 76% (1635/2150) of sepsis cases where sepsis screening was not initiated at triage. Finally, the results suggest that AI-based sepsis prediction in the ED may be able to significantly improve the time to antibiotics, which may offer opportunity for lifesaving intervention for patients. Notably, in addition to triage note vectors, the variables with the highest predictive importance were combinations of clinically relevant vital signs (time-of-triage model) and laboratory values, such as WBC count, creatinine, and lactic acid level (comprehensive model). These model characteristics, as well as the ability to map triage note word and subword relative contributions, indicate that the models may offer meaningfully explainable predictions to end users.

To our knowledge, this study is the largest to date to use NLP for sepsis prediction in the ED. We also demonstrated substantially improved accuracy compared to ML-based techniques in prior studies. The ability to incorporate triage notes into an ML model is advantageous for several reasons. First, natural language allows for a broad range of history and examination findings to be compressed into a short free-text note rather than innumerable variables in tabular form. Second, it allows experienced nurses to communicate an overall clinician impression that cannot always be captured by strictly quantitative inputs. In this study, free text from nursing triage notes was used to train a transformer model and was combined as input with other clinical data available at the time of initial triage, with the aim of predicting sepsis. Our findings demonstrate that NLP-based ML models can generate accurate predictions of sepsis at the time of triage and throughout an ED stay. Accordingly, the incorporation of free-text data into models that include data from clinical workups can produce a highly accurate prediction of sepsis.

Importance of Accurate Sepsis Prediction Tools

Existing sepsis alerting systems experience a number of performance difficulties. One of the most widely implemented sepsis detection systems across health systems has been shown to have limited performance due to low sensitivity and precision (33% and 2.4%, respectively). Low predictive performance hinders the clinical use of such systems, despite their aim being to prompt the initiation of lifesaving care. Further impacting their use are high rates of false positive alerts [12]. Increased rates of false positive alerts lead to lower trust among clinicians, alert fatigue and dismissal, and lower adoption [25]. Recently, the incorporation of natural language such as free-text notes into model inputs has been shown to be promising for accurately detecting sepsis as early as during the ED triage process [15].

Prior Studies

To our knowledge, this study is the largest to predict sepsis at the time of ED triage evaluation using NLP-based ML. Ivanov et al [15] reported high predictive performance for sepsis at ED triage with a smaller sample size in 2022. While both this study and Ivanov et al [15] present high sensitivity and specificity and remarkably increased performance compared to traditional screening tools for sepsis, there are important differences between the studies. Whereas Ivanov et al [15] included pediatric encounters, they were excluded in this study, since significantly ill patients of 18 years or younger of age are typically transferred to pediatric hospitals for admission and final diagnoses are unavailable. Accordingly, we excluded these encounters to avoid underestimation of sepsis in the pediatric population, which could have led to type I error with increased reliance on patient age as a predictive feature. A transformer model was also used for the NLP step, which can account for context and surrounding words.

Finally, our approach provides a method to present clinicians with understandable model decision explanations, including heat maps to indicate word importance and contribution to sepsis prediction. We present some examples of these heat maps here. It is important to note that the transformer architecture used in this study assigns meaning using full sentence context, capturing

combined subword and interword relationships, from negation to more complex interactions. As such, these heat maps can be instructive but offer a heavily simplified view of how the algorithm uses triage notes. Additionally, the triage note vectorization is only a part of our complete sepsis algorithm, which also considers additional clinical data throughout the ED encounter.

Limitations

There were several limitations in this study. First, physician-reviewed sepsis labels were only available for a subset of the data and had to be projected onto unlabeled encounters for training purposes using clinical signals. However, model performance was similar when evaluated on the secondary sepsis definition provided in the CDC hospital toolkit for adult sepsis surveillance. Second, the quality of the nursing triage notes is dependent on the clinical skill of the triage nurses, which could vary between EDs. Third, since the COVID-19 pandemic resulted in significant clinical and operational changes, it will be important to include such encounters in future prospective studies. Fourth, no pediatric patients were included, which would bias the model results toward an adult population. Fifth, in this

study, it was not possible to detect whether patients were immunocompromised. This is an important subgroup of patients to assess in future studies of ML-based sepsis prediction. Sixth, it was not possible in this study to stratify by causal organism of sepsis, which could affect performance characteristics. Finally, as this study was an investigation of NLP using triage notes, we excluded encounters having missing triage notes.

Conclusions

Using free-text and clinical data available at the time of initial ED triage from over 1 million patient encounters and across 4 hospital-based EDs, we demonstrated that NLP-based ML models are able to achieve high accuracy in predicting sepsis. The implication of these results is that AI-based clinical tools may substantially augment clinician abilities when clinical workup data are sparse, such as at the time of initial ED triage. Since sepsis mortality increases drastically with every passing hour and early clinical intervention is imperative to provide lifesaving treatment, AI-based tools using natural language data, such as free text available in nursing triage notes, may offer critical information to initiate treatment and prevent morbidity and mortality.

Conflicts of Interest

FB, NWS, SOF, and JDS are vice president of data science, machine learning research scientist, director of nursing, and cofounder and chief medical officer, respectively, at Vital Software, Inc, a company engaged in developing artificial intelligence clinical decision support products for the emergency department.

Multimedia Appendix 1

Examples of triage notes, subanalyses, and model explainability.

[[DOCX File, 1412 KB - ai_v3i1e49784_app1.docx](#)]

References

1. Hospital toolkit for adult sepsis surveillance. Centers for Disease Control and Prevention, Division of Healthcare Quality Promotion. 2018. URL: https://www.cdc.gov/sepsis/pdfs/Sepsis-Surveillance-Toolkit-Mar-2018_508.pdf [accessed 2023-04-01]
2. Fleischmann C, Scherag A, Adhikari NKJ, Hartog CS, Tsaganos T, Schlattmann P, et al. Assessment of global incidence and mortality of hospital-treated sepsis. Current estimates and limitations. *Am J Respir Crit Care Med* 2016;193(3):259-272 [FREE Full text] [doi: [10.1164/rccm.201504-0781OC](https://doi.org/10.1164/rccm.201504-0781OC)] [Medline: [26414292](https://pubmed.ncbi.nlm.nih.gov/26414292/)]
3. Rhee C, Dantes R, Epstein L, Murphy DJ, Seymour CW, Iwashyna TJ, et al. Incidence and trends of sepsis in US hospitals using clinical vs claims data, 2009-2014. *JAMA* 2017;318(13):1241-1249 [FREE Full text] [doi: [10.1001/jama.2017.13836](https://doi.org/10.1001/jama.2017.13836)] [Medline: [28903154](https://pubmed.ncbi.nlm.nih.gov/28903154/)]
4. Rudd KE, Johnson SC, Agesa KM, Shackelford KA, Tsoi D, Kievlan DR, et al. Global, regional, and national sepsis incidence and mortality, 1990-2017: analysis for the Global Burden of Disease Study. *Lancet* 2020;395(10219):200-211 [FREE Full text] [doi: [10.1016/S0140-6736\(19\)32989-7](https://doi.org/10.1016/S0140-6736(19)32989-7)] [Medline: [31954465](https://pubmed.ncbi.nlm.nih.gov/31954465/)]
5. Kashiouris MG, Zemore Z, Kimball Z, Stefanou C, Fowler AA, Fisher B, et al. Supply chain delays in antimicrobial administration after the initial clinician order and mortality in patients with sepsis. *Crit Care Med* 2019;47(10):1388-1395. [doi: [10.1097/CCM.0000000000003921](https://doi.org/10.1097/CCM.0000000000003921)] [Medline: [31343474](https://pubmed.ncbi.nlm.nih.gov/31343474/)]
6. Kumar A, Roberts D, Wood KE, Light B, Parrillo JE, Sharma S, et al. Duration of hypotension before initiation of effective antimicrobial therapy is the critical determinant of survival in human septic shock. *Crit Care Med* 2006;34(6):1589-1596. [doi: [10.1097/01.CCM.0000217961.75225.E9](https://doi.org/10.1097/01.CCM.0000217961.75225.E9)] [Medline: [16625125](https://pubmed.ncbi.nlm.nih.gov/16625125/)]
7. Angus DC. The lingering consequences of sepsis: a hidden public health disaster? *JAMA* 2010;304(16):1833-1834. [doi: [10.1001/jama.2010.1546](https://doi.org/10.1001/jama.2010.1546)] [Medline: [20978262](https://pubmed.ncbi.nlm.nih.gov/20978262/)]
8. Liang L, Moore B, Soni A. National inpatient hospital costs: the most expensive conditions by payer, 2017. Agency for Healthcare Research and Quality. 2020. URL: <https://www.hcup-us.ahrq.gov/reports/statbriefs/sb261-Most-Expensive-Hospital-Conditions-2017.jsp> [accessed 2023-02-05]

9. Dellinger RP, Levy MM, Rhodes A, Annane D, Gerlach H, Opal SM, et al. Surviving sepsis campaign: international guidelines for management of severe sepsis and septic shock: 2012. *Crit Care Med* 2013;41(2):580-637 [FREE Full text] [doi: [10.1097/CCM.0b013e31827e83af](https://doi.org/10.1097/CCM.0b013e31827e83af)] [Medline: [23353941](https://pubmed.ncbi.nlm.nih.gov/23353941/)]
10. Jaimes F, Garcés J, Cuervo J, Ramírez F, Ramírez J, Vargas A, et al. The systemic inflammatory response syndrome (SIRS) to identify infected patients in the emergency room. *Intensive Care Med* 2003;29(8):1368-1371. [doi: [10.1007/s00134-003-1874-0](https://doi.org/10.1007/s00134-003-1874-0)] [Medline: [12830377](https://pubmed.ncbi.nlm.nih.gov/12830377/)]
11. Perman SM, Mikkelsen ME, Goyal M, Ginde A, Bhardwaj A, Drumheller B, et al. The sensitivity of qSOFA calculated at triage and during emergency department treatment to rapidly identify sepsis patients. *Sci Rep* 2020;10(1):20395 [FREE Full text] [doi: [10.1038/s41598-020-77438-8](https://doi.org/10.1038/s41598-020-77438-8)] [Medline: [33230117](https://pubmed.ncbi.nlm.nih.gov/33230117/)]
12. Wong A, Oates E, Donnelly JP, Krumm A, McCullough J, DeTroyer-Cooley O, et al. External validation of a widely implemented proprietary sepsis prediction model in hospitalized patients. *JAMA Intern Med* 2021;181(8):1065-1070 [FREE Full text] [doi: [10.1001/jamainternmed.2021.2626](https://doi.org/10.1001/jamainternmed.2021.2626)] [Medline: [34152373](https://pubmed.ncbi.nlm.nih.gov/34152373/)]
13. Bennett T, Russell S, King J, Schilling L, Voong C, Rogers N, et al. Accuracy of the epic sepsis prediction model in a regional health system. *ArXiv. Preprint posted online on February 19, 2019* 2019 [FREE Full text]
14. Shashikumar SP, Josef CS, Sharma A, Nemati S. DeepAISE—an interpretable and recurrent neural survival model for early prediction of sepsis. *Artif Intell Med* 2021;113:102036 [FREE Full text] [doi: [10.1016/j.artmed.2021.102036](https://doi.org/10.1016/j.artmed.2021.102036)] [Medline: [33685592](https://pubmed.ncbi.nlm.nih.gov/33685592/)]
15. Ivanov O, Molander K, Dunne R, Liu S, Masek K, Lewis E, et al. Accurate detection of sepsis during emergency department triage using machine learning. *ArXiv. Preprint posted online on April 15, 2022* 2023 [FREE Full text]
16. Sterling NW, Brann F, Patzer RE, Di M, Koebbe M, Burke M, et al. Prediction of emergency department resource requirements during triage: an application of current natural language processing techniques. *J Am Coll Emerg Physicians Open* 2020;1(6):1676-1683 [FREE Full text] [doi: [10.1002/emp2.12253](https://doi.org/10.1002/emp2.12253)] [Medline: [33392576](https://pubmed.ncbi.nlm.nih.gov/33392576/)]
17. Sterling NW, Patzer RE, Di M, Schragger JD. Prediction of emergency department patient disposition based on natural language processing of triage notes. *Int J Med Inform* 2019;129:184-188. [doi: [10.1016/j.ijmedinf.2019.06.008](https://doi.org/10.1016/j.ijmedinf.2019.06.008)] [Medline: [31445253](https://pubmed.ncbi.nlm.nih.gov/31445253/)]
18. Zhang X, Kim J, Patzer RE, Pitts SR, Patzer A, Schragger JD. Prediction of emergency department hospital admission based on natural language processing and neural networks. *Methods Inf Med* 2017;56(5):377-389. [doi: [10.3414/ME17-01-0024](https://doi.org/10.3414/ME17-01-0024)] [Medline: [28816338](https://pubmed.ncbi.nlm.nih.gov/28816338/)]
19. Barrett ML, Owens PL, Roemer M. Changes in emergency department visits in the initial period of the COVID-19 pandemic (april–december 2020), 29 states. In: *Healthcare Cost and Utilization Project (HCUP) Statistical Briefs. Statistical Brief #298*. Rockville, MD: Agency for Healthcare Research and Quality (US); 2006.
20. Boserup B, McKenney M, Elkbuli A. The impact of the COVID-19 pandemic on emergency department visits and patient safety in the United States. *Am J Emerg Med* 2020;38(9):1732-1736 [FREE Full text] [doi: [10.1016/j.ajem.2020.06.007](https://doi.org/10.1016/j.ajem.2020.06.007)] [Medline: [32738468](https://pubmed.ncbi.nlm.nih.gov/32738468/)]
21. Devlin J, Chang M, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. 2019 Presented at: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*; June 2-7, 2019; Minneapolis, MN, USA.
22. Sanh V, Debut L, Chaumond J, Wolf T. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *ArXiv. Preprint posted online on October 2, 2019* 2019 [FREE Full text]
23. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. 2016 Presented at: *KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*; August 13-17, 2016; San Francisco, CA, USA p. 785-794. [doi: [10.1145/2939672.2939785](https://doi.org/10.1145/2939672.2939785)]
24. Sundararajan M, Taly A, Yan Q. Axiomatic attribution for deep networks. 2017 Presented at: *ICML'17: Proceedings of the 34th International Conference on Machine Learning—Volume 70*; August 6-11, 2017; Sydney, New South Wales, Australia p. 3319-3328.
25. Henry KE, Adams R, Parent C, Soleimani H, Sridharan A, Johnson L, et al. Factors driving provider adoption of the TREWS machine learning-based early warning system and its effects on sepsis treatment timing. *Nat Med* 2022;28(7):1447-1454. [doi: [10.1038/s41591-022-01895-z](https://doi.org/10.1038/s41591-022-01895-z)] [Medline: [35864251](https://pubmed.ncbi.nlm.nih.gov/35864251/)]

Abbreviations

- AI:** artificial intelligence
- AUC:** area under the receiver operating characteristic curve
- BERT:** Bidirectional Encoder Representation From Transformers
- CDC:** Centers for Disease Control and Prevention
- ED:** emergency department
- ICU:** intensive care unit
- ML:** machine learning
- NLP:** natural language processing

WBC: white blood cell

XGBoost: Extreme Gradient Boosting

Edited by K El Emam, B Malin; submitted 08.06.23; peer-reviewed by L Prunelli; comments to author 10.07.23; revised version received 15.08.23; accepted 16.12.23; published 25.01.24.

Please cite as:

Brann F, Sterling NW, Frisch SO, Schrage JD

Sepsis Prediction at Emergency Department Triage Using Natural Language Processing: Retrospective Cohort Study

JMIR AI 2024;3:e49784

URL: <https://ai.jmir.org/2024/1/e49784>

doi: [10.2196/49784](https://doi.org/10.2196/49784)

PMID: [38875594](https://pubmed.ncbi.nlm.nih.gov/38875594/)

©Felix Brann, Nicholas William Sterling, Stephanie O Frisch, Justin D Schrage. Originally published in JMIR AI (<https://ai.jmir.org>), 25.01.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR AI, is properly cited. The complete bibliographic information, a link to the original publication on <https://www.ai.jmir.org/>, as well as this copyright and license information must be included.

Original Paper

Learning From International Comparators of National Medical Imaging Initiatives for AI Development: Multiphase Qualitative Study

Kassandra Karpathakis¹, BA, BSc, MPH; Emma Pencheon², BA, MBBS, MSc; Dominic Cushnan³, MBA

¹Decimal.health, Boston, MA, United States

²Foreign, Commonwealth and Development Office, UK Government, London, United Kingdom

³NHS England, London, United Kingdom

Corresponding Author:

Kassandra Karpathakis, BA, BSc, MPH

Decimal.health

50 Milk Street

Boston, MA, 02109

United States

Phone: 1 6086285988

Email: kass.karpathakis@gmail.com

Abstract

Background: The COVID-19 pandemic drove investment and research into medical imaging platforms to provide data to create artificial intelligence (AI) algorithms for the management of patients with COVID-19. Building on the success of England's National COVID-19 Chest Imaging Database, the national digital policy body (NHSX) sought to create a generalized national medical imaging platform for the development, validation, and deployment of algorithms.

Objective: This study aims to understand international use cases of medical imaging platforms for the development and implementation of algorithms to inform the creation of England's national imaging platform.

Methods: The National Health Service (NHS) AI Lab Policy and Strategy Team adopted a multiphased approach: (1) identification and prioritization of national AI imaging platforms; (2) Political, Economic, Social, Technological, Legal, and Environmental (PESTLE) factor analysis deep dive into national AI imaging platforms; (3) semistructured interviews with key stakeholders; (4) workshop on emerging themes and insights with the internal NHSX team; and (5) formulation of policy recommendations.

Results: International use cases of national AI imaging platforms (n=7) were prioritized for PESTLE factor analysis. Stakeholders (n=13) from the international use cases were interviewed. Themes (n=8) from the semistructured interviews, including interview quotes, were analyzed with workshop participants (n=5). The outputs of the deep dives, interviews, and workshop were synthesized thematically into 8 categories with 17 subcategories. On the basis of the insights from the international use cases, policy recommendations (n=12) were developed to support the NHS AI Lab in the design and development of the English national medical imaging platform.

Conclusions: The creation of AI algorithms supporting technology and infrastructure such as platforms often occurs in isolation within countries, let alone between countries. This novel policy research project sought to bridge the gap by learning from the challenges, successes, and experience of England's international counterparts. Policy recommendations based on international learnings focused on the demonstrable benefits of the platform to secure sustainable funding, validation of algorithms and infrastructure to support in situ deployment, and creating wraparound tools for nontechnical participants such as clinicians to engage with algorithm creation. As health care organizations increasingly adopt technological solutions, policy makers have a responsibility to ensure that initiatives are informed by learnings from both national and international initiatives as well as disseminating the outcomes of their work.

(JMIR AI 2024;3:e51168) doi:[10.2196/51168](https://doi.org/10.2196/51168)

KEYWORDS

digital health; mobile health; mHealth; medical imaging; artificial intelligence; health policy

Introduction

Background

Medical imaging has been identified by many governments as an especially promising application for artificial intelligence (AI) in clinical practice with the potential to enhance disease screening, improve care outcomes, and reduce costs [1-5]. Optimizing AI capabilities requires aggregating and streamlining access to medical imaging data for machine learning (ML) model training and validation and contextualized mechanisms for deployment in clinical workflows.

During England's National Health Service (NHS) response to the COVID-19 pandemic, the digital health agency (NHSX) created the National COVID-19 Chest Imaging Database (NCCID). The NCCID is a "centralized UK database containing chest X-rays (CXR), Computer Tomography (CT) and Magnetic Resonance Images (MRI) from hospital patients" with COVID-19 [6,7]. It was established to develop, validate, and deploy AI and ML models for supporting the management of patients with severe COVID-19. The creation of the NCCID highlighted the merits and challenges of a centralized approach for collating national imaging data [7].

The NCCID led to a proposal for a generalized national imaging platform for the development, validation, and deployment of AI and ML models in medical imaging. This platform was envisaged to have three technical functions:

1. A data pipeline to facilitate the collection of data nationally
2. A trusted research environment (TRE) to provide access to national data to build and validate new AI and ML products
3. A deployment platform to act as an "app store" for the most up-to-date AI and ML models for users in health care facilities

To support the safe, ethical, and effective creation and deployment of a national imaging platform, the NHS AI Lab developed complementary policy and regulatory initiatives, including a cross-regulatory service to guide developers through the regulation of their AI products [8], understanding of public attitudes toward sharing health data for AI development, and an Algorithmic Impact Assessment tool to identify potential societal impacts of AI products [9].

Beyond understanding the policy and infrastructural requirements, it is important to assess the strengths and weaknesses of such a national approach to produce AI and ML models for imaging that can be deployed in clinical workflows. To make such an assessment, the NHS AI Lab analyzed international efforts to build similar medical imaging platforms in both private and public organizations, some of which were associated with national efforts to diagnose and manage patients with COVID-19. The NHS AI Lab used the outputs of the research to understand the approaches taken and lessons learned and inform the design of England's national imaging platform.

Objectives

We sought to identify and understand international use cases of and proposals for medical imaging platforms to streamline the innovation-to-deployment journey for health AI models in

imaging. We aimed to understand how imaging for AI efforts were structured, identify the constituent parts of the initiatives (eg, technical aspects, users and marketplace, and commercialization), and understand the implications of government policy and regulation. We used this analysis of international use cases to formulate policy recommendations for England's nascent national AI imaging platform.

Methods

Overview

This research was conducted by NHSX, the former digital health agency and technology policy arm of NHS England. NHSX was merged into the NHS England transformation directorate in 2022. The Strategy and Policy Team at the NHS AI Lab, which was embedded inside NHSX, led and completed the study. This project was conducted between September 2020 and March 2021.

Phase 1: Identification and Prioritization of National AI Imaging Platforms

We conducted a preliminary scan to identify efforts to create national AI imaging platforms in other countries that the NHS AI Lab could analyze in depth.

As the United Kingdom was poised to lead the G7 in 2021, we started with fellow G7 countries: Canada, France, Germany, Italy, Japan, and the United States of America. We then scanned non-G7 countries known within digital health policy circles for their digital health approaches or that had previously responded to an NHSX survey on the use of AI by Global Digital Health Partnership (GDHP) member countries [10]: Australia, Brazil, China, Estonia, Hong Kong, India, Republic of Korea, Rwanda, Singapore, Sweden, Uganda, and Uruguay. Finally, we scanned initiatives in multilateral collaborations (World Health Organization, International Telecommunication Union, and the GDHP) and major private organizations (eg, GE Healthcare and Google).

National AI imaging initiatives were identified by 2 researchers (Abhishek Mishra and EP) through (1) a targeted Google search for each country using [country] and the keywords *AI medical imaging platform*, *medical imaging data*, *medical AI platform*, *AI radiology*, or *COVID-19 medical image AI*; (2) a targeted Google search for multilateral collaborations and major private organizations using [name of organization] and the keywords *AI medical imaging platform*, *medical imaging data*, *medical AI platform*, *AI radiology*, or *COVID-19 medical image AI*; and (3) a general search on Google, Google Scholar, Twitter, and One HealthTech using the keywords *medical imaging AI platform*, *medical imaging platform*, *national medical imaging AI platform*, or *medical imaging AI marketplace*. For each search, the first 5 pages of the results were scanned owing to time and resource limitations.

We scored each initiative in comparison with the United Kingdom's context to prioritize some for the deeper dive in phase 2. Each of the following criteria (n=4) was scored from similar (score=3) to not similar (score=1); initiatives with the

highest total score were deemed most similar to that of the United Kingdom:

1. Similarity of the medical imaging platform to the United Kingdom's proposed initiative: medical imaging data only versus additional health data, TRE built on top of data to allow for model development, data consolidated in a centralized location or alternative approaches such as federated learning, and parallel building of deployment platform.
2. Size of market: using the country population as a proxy — ≥ 50 million, 10 to 50 million, and 0 to 10 million.
3. Future trade importance to the United Kingdom: priority markets identified by the NHS Director of AI based on track record of digital health initiatives (note that, at the time of the study, the United Kingdom was the Chair of the G7, and there was strong political interest in the potential for health AI to bolster the United Kingdom's trade agenda).
4. Regulatory and ecosystem similarity to that of the United Kingdom based on the following: provincial versus national digital health organization, single-payer versus multipayer system, and regulatory approach to AI.

Phase 2: Deep Dive Into National AI Imaging Platforms

For the prioritized initiatives, we conducted a deep dive using the Political, Economic, Social, Technological, Legal, and Environmental (PESTLE) factors framework. PESTLE is a common tool used in policy analysis to gain an overview of an industry [11].

The aims of the deep dive were to (1) identify reliable and robust information to inform the understanding of the international use case; (2) identify hypotheses, gaps, and insights on the AI imaging initiatives for validation during stakeholder interviews; and (3) inform the creation of a deductive framework for the analysis of semistructured interviews. We also identified stakeholders leading AI initiatives to approach for the semistructured interviews in phase 3.

Phase 3: Semistructured Interviews

Semistructured interviews were conducted to understand each prioritized initiative (eg, data used and intended users); its social and political context (eg, regulatory landscape, stakeholders, and public trust), data handling (eg, data and privacy laws), funding sources, and commercialization; and the lessons learned during its development. The discussion guide ([Multimedia Appendix 1](#)) was tailored to each country's unique imaging platform, including the validation of any gaps or insights identified in phase 2.

The interviews were conducted by one principal researcher (KK) with one supporting researcher (EP). Informed consent was obtained from interview participants, and they approved the selected quotes for publication. The interviews lasted up to 1 hour and were audio recorded, and detailed notes were taken. Transcription and translation services were provided by an independent agency. Only one country (Singapore) required the use of translation services to conduct the interview. All other interviews were conducted in English. Both the detailed notes and transcripts from the interviews were analyzed.

The interviews were analyzed using a deductive framework with codes identified from the desk research deep dives ([Multimedia Appendix 2](#)). In total, 2 researchers (KK and EP) analyzed each interview independently and compared their coding. Intercoder reliability (ICR) was calculated to assess the reliability of the coding protocol and thematic analysis. ICR was calculated by comparing the level of agreement and disagreement across the coding for 5 pages per transcript [12].

Phase 4: Workshop With NHS AI Lab National Imaging Platform Team

A workshop was conducted with the NHS AI Lab national AI imaging platform team members who were conducting the discovery phase [13]. The workshop aims were to (1) establish top areas of interest from the perspective of the discovery team, (2) explore why these areas are important to the team, and (3) stimulate the discovery teams' interest in applying the lessons learned from other countries.

The workshop was facilitated by one principal researcher (KK) with one supporting researcher (EP). The workshop lasted 90 minutes, and audio recordings and detailed notes were taken. Participants (n=5) used the web-based *Padlet* and *Jamboard* (Google) post-it and "like" functionalities. If required, the researchers noted the participants' points on their behalf. The workshop audio was transcribed and analyzed.

An overview of the initiatives (n=6) from phase 2 and phase 3 was provided to the attendees using *Jamboard*. The countries were treated as individual case studies rather than grouped together because of the large degree of heterogeneity between the countries.

A total of 8 themes from the deductive framework were used to guide the workshop: purpose; users; organizational; commercialization; data; incentives; building trust; and law, policy, and regulation. Quotes from the semistructured interviews with stakeholders (phase 3) from each initiative were mapped to the 8 themes for discussion at the workshop.

The nominal group technique was used to identify priority quotes and insights [14]. Participants were asked to vote on the quotes that resonated or were of interest to them using *Padlet*'s "like" functionality. Each participant had 6 votes per initiative. Voting indicated the discovery team's priorities and fueled discussions.

The outputs of the deep dives, interviews, and workshop were synthesized thematically into 8 categories with 17 subcategories. The analysis was inspired by a user-centered design insight format [15], which states the context and background, explains the learning, explains the root cause (the why), and explains the motivation behind why the learning has occurred and the ramifications for the NHS AI Lab's proposed national medical AI imaging initiative.

Phase 5: Formulating Recommendations

The researchers (KK and EP) jointly synthesized all the data gathered from phase 3 to phase 4 to formulate recommendations for the NHS AI Lab national AI imaging initiative. This involved drawing out themes based on the original thematic framework,

identifying learnings pertinent to the United Kingdom, and framing the resulting insights into actionable recommendations.

Final recommendations were presented to the Head of AI Imaging and Director of AI at the NHS AI Lab for consideration. The Head of AI Imaging and the national AI imaging discovery team selected the recommendations that were relevant and actionable for the discovery and future phases of the project. The research team was not privy to this selection.

Ethical Considerations

Internal and external stakeholders were consulted during this policy research and development. Informed consent was obtained from interview and workshop participants. Per NHSX's standard practice, independent ethical review was not required for this research informing policy as it poses negligible risk.

Results

Phase 1: Identified National AI Medical Imaging Platforms

Numerous initiatives (n=34) were identified from preliminary scanning. Most initiatives were country based (21/34, 62%), and the remainder were from major private organizations (10/34, 29%) or multinational organizations (3/34, 9%). Some of the initiatives (7/34, 21%) were prioritized for a deep dive: (1) Digital Health and Discovery Platform (DHDP; Canada), (2) national medical image database (China), (3) Hospital Authority Data Collaboration Laboratory (HADCL; Hong Kong), (4) Research Center for Medical Big Data (Japan), (5) AI Medical Imaging Platform (Singapore), (6) Analytic Imaging Diagnostics Arena (AIDA; Sweden), and (7) Medical Imaging and Data Resource Center (MIDRC; United States).

Phases 2 and 3: Overview of Prioritized National AI Imaging Platforms

In the following sections, we provide a brief overview of each initiative. [Multimedia Appendix 3 \[16-44\]](#) provides a detailed overview of each country's initiative complemented with findings from the PESTLE analysis and semistructured interviews.

Canada: DHDP

This pan-Canadian initiative was set up to create a nationwide framework to digitally enable research that advances next-generation precision medicine technologies with an emphasis on cancer and improving health outcomes for patients. The DHDP comprises >90 consortium partners spanning academia and the private sector. The initiative focused on numerous types of medical data rather than solely on medical imaging [45] and undertook novel research in federated learning technologies that reflected Canada's stringent attitudes toward data privacy and sharing.

China: National Medical Image Database

In September 2020, plans were announced for the creation of a standardized national medical image database. The Chinese national medical image database was approved by the National Health Commission [19] to enable hospitals to share patient information and medical images and support the training and

development of AI technology for health care. At the time of the study, it was unclear what technology stack the Chinese national imaging database would use and how the initiative would overcome issues of data digitization, cybersecurity, and commercialization.

Hong Kong: HADCL

The HADCL was established to support the formulation of health care policies, facilitate biotechnological research, and improve clinical and health care services. The HADCL is the flexible and interactive data-sharing channel of Hong Kong's Hospital Authority, with a growing focus on the development of AI and ML algorithms. It is a full-service offering encouraging researchers to partake in collaborative health data projects in a controlled environment using the Hospital Authority's extensive, longitudinal data [46,47].

Japan: Research Center for Medical Big Data

Japan's Research Center for Medical Big Data is a platform for AI technology research and development, including a cloud-based platform for hosting medical imaging big data and analyzing medical images. As of 2019, the platform contained >10 million medical images, with participation from at least 60 hospitals. In line with policy at the time of the study, the platform's primary user base was academia, and projects were for research purposes only.

Singapore: AI-Enabled Medical Imaging Platform

In October 2020, the Integrated Health Information System health laboratory issued a call for collaboration between partners to cocreate an "AI-enabled Medical Imaging Platform" aimed at operationalizing and exploring AI models and applications for medical imaging. The platform will be open and vendor neutral, thereby enabling the deployment of AI models and products from different sources to assist with clinicians' work.

Sweden: AIDA

AIDA is a dedicated initiative for research and innovation in AI and medical image analysis in Sweden. The initiative brings together academia, health care, and industry to translate innovation into AI-based decision support solutions for imaging diagnostics. The previous mandated creation of national registries containing >5 TB of health data provided the foundation for the AIDA initiative.

United States: MIDRC

The MIDRC is a multi-institutional initiative established in response to the COVID-19 pandemic. The aim was to foster ML innovation through the sharing of imaging and associated clinical data regarding COVID-19 [48]. At the time of the study, agreements for sharing relevant medical imaging data were in the process of being signed with several sites, but no data were being hosted on the platform.

Phases 3 and 4: Derived Themes and Insights

Stakeholders (n=16) representing 7 initiatives were approached for interviews. Stakeholders (n=13) from 6 initiatives accepted the interview invitations (13/16, 81% acceptance rate). The stakeholders from participating countries were 38% (5/13) from Canada, 8% (1/13) from Hong Kong, 23% (3/13) from Japan,

8% (1/13) from Singapore, 8% (1/13) from Sweden, and 15% (2/13) from the United States. Stakeholders from China (3/16, 19%) did not respond to the request for an interview.

For the interview coding, the ICR between the researchers (KK and EP) was calculated to be 0.41, indicating moderate reliability [12,49]. The outputs of the deep dives, interviews, and workshop were synthesized thematically into categories (n=8) with subcategories (n=17).

[Multimedia Appendix 4](#) presents the categories, subcategories, and corresponding thematic synthesis within each of the other countries' initiative including key insights, quotes, and learnings.

Phase 5: Recommendations

Overview

We provided 12 recommendations for the NHS AI Lab's proposed national AI imaging platform. Each recommendation is grounded in the themes and insights from phase 2 to phase 4 (see [Multimedia Appendix 4](#)). The corresponding themes for each recommendation are also provided.

Narrative

Recommendation 1: The NHS AI Lab develop a purposeful narrative of why and how a national medical imaging initiative is necessary, outlining what health needs it will meet and supporting this with demonstration of its benefit and potential

Developing a strong value proposition should be married with demonstrable benefit. The narrative should be cross-cutting, speaking not only to purpose but also to trust and incentives, with transparency regarding the drivers of the initiative. Previous work by the NHS AI Lab on behalf of the GDHP has also argued that countries should take a "needs based" approach to AI-driven technology development to create both maximal benefit on health outcomes and foster buy-in and support from stakeholders and the public [47,50].

A purposeful narrative for the NHS AI Lab's national medical imaging initiative will support interdisciplinary collaboration and ensure long-term political, financial, and social support for the initiative based on a clear understanding of its importance and utility to the health system. An important aspect of this narrative is to reference the value of the initiative as a social or public good that creates public value [51].

The corresponding themes for this recommendation are (A) demonstrable benefit of the initiative, (B) health system needs as the primary driver, (C) community and shared purpose, and (O) transparency and communication. transparency and communication.

Recommendation 2: The NHS AI Lab moves away from the language of "platform" to talking about the national medical imaging initiative as an "initiative" and community space for growing the United Kingdom's understanding and ability to use AI in medical imaging

The United Kingdom's national medical imaging "initiative" should be carefully framed, using language that reflects what is offered and conveys mindset and purpose. The connotations

of "national" in the initiative name given the involvement (or lack thereof) of the Devolved Administrations (DAs) should be considered. In addition, the NHS AI Lab should develop an approach for involving the DAs.

The corresponding themes for this recommendation are (C) community and shared purpose and (D) embracing and enabling the central role of health care professionals.

Users and Service Offering

Recommendation 3: The NHS AI Lab develops wraparound services to maximize engagement and capitalize on the expertise of varied users; by removing the need to technically upskill in AI development while also providing opportunities for users to do so if they wish, the initiative can broaden participation and avoid disincentivizing users with different and valuable areas of specialty

The NHS AI Lab should invest in wraparound services, specifically offering tools and professional technical skills that are tailored to fill a gap that users, such as health care professionals, have when it comes to developing AI. It appears from international comparators that the main draw and success has not been the platform itself but the supportive services to enable users to engage, collaborate, and develop AI-driven technologies regardless of their technical expertise. Examples include but are not limited to clinical fellowships on health data, networking or pairing clinicians with data scientists, training courses on what is AI and how to develop models, and low-code and no-code AI model development tools. The NHS AI Lab should explore opportunities to build these wraparound services from existing programs in the digital health ecosystem.

The corresponding themes for this recommendation are (D) embracing and enabling the central role of health care professionals, (E) recognizing that users are not discrete groups, and (F) importance of wraparound services.

Recommendation 4: The NHS AI Lab continues to embrace interdisciplinary work while designing, developing, and implementing the national medical imaging initiative; the inherent tensions and perspectives between disciplines are needed to deliver on health system needs

Interdisciplinary work is central to harnessing the breadth of expertise required to build and sustain an initiative that truly addresses health system needs. This means embracing the central role of health care professionals and ensuring the participation of people who have a system view of health and social care, as well as those with frontline experience who will be the ultimate end users of any AI products developed on the platform. Prioritizing user-centered design and health care professionals' experience means that technical expertise must take an important facilitative and instructive role to both guide and learn from health care professionals about how to leverage AI-driven technologies in the health system. By facilitating interdisciplinary work, radiologists' expertise can be applied to shore up the quality and appropriateness of the imaging data used. We recommend that active steps be taken to foster collaborative working relationships across disciplines drawing

on lessons for interdisciplinary collaboration outlined by Blandford et al [52] and on the examples of activities run in Sweden and Japan.

The corresponding themes for this recommendation are (B) health system needs as the primary driver for AI development, (D) embracing and enabling the central role of health care professionals, and (E) recognizing that users are not discrete groups.

Sustainability and Future-Proofing

Recommendation 5: The NHS AI Lab consider the financial sustainability of the national medical imaging initiative from the outset and how this maps to the proposed commercial model

All the international comparators who did not have a clear commercial model raised concerns about financial sustainability. It is worth bearing in mind that demonstrable benefit does not guarantee enduring government support with respect to funding. We recommend that the NHS AI Lab national medical imaging initiative considers how the work will be sustained beyond current funding and ensures that options for commercialization are not excluded by virtue of how the initiative is designed (ie, data-sharing arrangements that preclude commercialization). For the NHS AI Lab's national medical imaging initiative to have longevity, it is important to keep as many commercial options on the table as possible, including generating revenue from certain aspects of the initiative and exploring public-private partnerships. This could include providing data subsets to fulfill specific needs, such as validation, that can be commercialized as a distinct offering.

The corresponding themes for this recommendation are (I) ensure financial sustainability, (J) differing or absent commercial models, and (L) subsetting data offerings.

Recommendation 6: The NHS AI Lab continues to explore different commercial models for the national medical imaging initiative with a focus on how it might commercialize aspects of the initiative rather than taking an all-or-none approach

Commercialization is likely necessary to ensure the financial sustainability of the initiative. Commercial models were an afterthought for many international comparators, who conveyed the sense that commercialization was viewed as being in opposition to the public good. We recommend thinking about commercial options early on, not only from a practical perspective of building the initiative with this in mind but also to construct a narrative that can interweave commercialization and private sector involvement with the public good. The NHS AI Lab should continue working with internal teams (ie, the NHSX Centre for Improving Data Collaboration) to ensure that the NHS gains fair value for the public from commercial arrangements.

The corresponding themes for this recommendation are (I) ensure financial sustainability, (J) differing or absent commercial models, and (N) a focus on public and social good.

Recommendation 7: The NHS AI Lab explore and potentially adopt some of the future-proofing mechanisms used by international comparators

Sweden and the United States exemplified ways to future-proof data-sharing mechanisms, including specific clauses in data-sharing agreements that granted them the power to revoke data access or extend it to future offerings. This is important for safeguarding against issues further down the road and streamlining the process of setting up data-sharing agreements. Sweden was cognizant that currently, anonymized data might become reidentifiable with advances in data analysis and wanted to mitigate this risk from the outset through the ability to revoke access at any time. We also recommend that, if and where possible, the initiative infrastructure is future-proofed and reusable so that it will be fit for purpose in years to come and offer benefits to other similar initiatives.

The corresponding themes for this recommendation are (M) future-proofing mechanisms for data sharing and (N) a focus on public and social good.

Recommendation 8: The NHS AI Lab balances the need to deliver at pace with the up-front investment of time and effort required to ensure that the resulting initiative is sustainable and future-proofed

A variety of pressures to deliver at pace were identified by international colleagues, which at times nudged countries toward "kicking the can down the road" when it came to thorny challenges such as commercialization. Although a certain level of pace is necessary to demonstrate benefit and garner support, this should be tempered to ensure an up-front investment of time and effort that delivers sustainable returns.

The corresponding themes for this recommendation are (A) demonstrable benefit of the national medical imaging initiative and (G) tempering the pace of development.

Recommendation 9: The NHS AI Lab consider under what conditions it would be acceptable and feasible to move beyond human-in-the-loop approaches in the national medical imaging initiative's resultant AI-driven technologies

All countries maintained the need for a human to be "in the loop" to ensure the safety, accountability, and acceptability of AI development and products. *Human-in-the-loop* refers to models that require human interaction, whereby human oversight can intervene and determine the outcome of a process or event. However, there is an undertone that moving beyond human-in-the-loop approaches is the future state of AI-driven technology in health and care (in some conditions, not yet defined). We recommend that the NHS AI Lab start considering not only the safety and accountability of systems without humans and when this would be deemed appropriate but also the public perception of not having unique or individualized care.

The corresponding themes for this recommendation are (K) common and continuing data challenges, (O) transparency and communication, and (P) keeping humans in the loop.

Recommendation 10: The NHS AI Lab accounts for the environmental impact of the national medical imaging initiative and establishes how it aligns with a sustainable health and social care system

No international comparators had considered the environmental impact of their initiative or how they were positioned in relation to delivering a sustainable health and care system. This presents an opportunity for the United Kingdom to lead in this domain considering the health system needs not only for now but also for the future. We recommend that the NHS AI Lab develop an understanding of how the national medical imaging initiative could affect both positively and negatively an economically and environmentally sustainable health system. This is an important element of future-proofing the work and ensuring that it is fit for purpose in the coming decades (note: the NHS AI Lab strategy team has started considering how AI could contribute to the NHS goal of reaching net zero by 2045 and to an environmentally sustainable health and care system [53]).

The corresponding themes for this recommendation are (B) health system needs as the primary driver and (N) a focus on public and social good.

Policy and Regulation

Recommendation 11: The NHS AI Lab leverage its privileged position as the guiding health technology organization within both the civil service and the NHS to continue advocating and driving policy and regulatory change; the United Kingdom's national medical imaging initiative is a tangible use case for uncovering the issues and providing examples of how they could be solved

All countries recognized that their current policies and regulations were not fit for the purpose of AI development and implementation in clinical settings. There was a range of mindsets regarding how to balance operating within constraints and advocating to change them. The NHS AI Lab is uniquely positioned within the government to drive the necessary changes in the United Kingdom making use of existing collaborations with regulatory bodies and DAs. We recommend that the national medical imaging initiative, with clearly articulated and demonstrable benefits to the health system, be used as evidence for this advocacy work.

The corresponding themes for this recommendation are (H) building on existing infrastructure and resources and (Q) advocating for policy, regulatory, and legal frameworks that are fit for purpose.

Recommendation 12: The NHS AI Lab leverage the work already undertaken in validation of AI models as a unique selling point for the United Kingdom's national medical imaging initiative

No international comparators had progressed to the deployment and widespread adoption of AI-driven technologies developed through their initiatives. One of the bottlenecks for this is a clear validation process, an area in which the NHS AI Lab is well placed to take the lead given the existing work that has been done in this domain. We recommend that this is capitalized on as a unique selling point for the national medical imaging

initiative to demonstrate an innovation funnel that runs smoothly through to the deployment of assured technologies.

The corresponding themes for this recommendation are (H) building on existing infrastructure and resources and (Q) advocating for policy, regulatory, and legal frameworks that are fit for purpose.

Discussion

Principal Findings

The NHS AI Lab sought to learn from countries developing medical imaging platforms to streamline the innovation-to-deployment journey for AI and ML algorithms for medical imaging. The research team conducted secondary and primary research with use cases from multiple countries to develop a deep understanding of the approaches for structuring a medical imaging platform program, how to set up supportive policy and regulatory initiatives, and form relationships with international stakeholders.

In addition to providing 12 recommendations for the NHS AI Lab to implement, the research team identified five areas in which the NHS AI Lab could offer a unique value proposition:

1. Galvanizing the already operating proof of concept, the NCCID program, to demonstrate benefit and secure stable United Kingdom government funding and support.
2. Within the new medical imaging platform, build in the ability to validate AI and ML algorithms as well as deploy them in health care settings. Only a few international initiatives built in the ability to validate algorithms and create a deployment pipeline, which is crucial for ensuring the effectiveness of algorithms during implementation.
3. Create wraparound offerings tailored to researchers, developers, and private companies operating in the United Kingdom. This may include tools to facilitate the creation of algorithms, training and workshops for upskilling, computational power, legal and regulatory support, and demand signaling for areas of clinical specialty in which there is high demand for AI and ML development.
4. Consider the environmental impact and sustainability of the medical imaging platform and the resultant carbon output from the outset.
5. Publicly demonstrate that the NHS AI Lab has incorporated collaborative international learnings and best practices.

Strengths

The primary strength of the project was the NHS AI Lab's openness to learning from other countries. Throughout our engagement with selected countries (Canada, Hong Kong, Japan, Singapore, Sweden, and the United States), we established that no other initiative had conducted international landscaping to inform strategy and implementation. Our work highlights the benefit of not reinventing the wheel in health AI initiatives but reaching out to build on the experience and expertise of others.

Second, the internal discovery team responsible for designing and building the NHS AI Lab's medical imaging platform was engaged throughout the delivery of this project. Their engagement culminated in the workshop to elicit feedback and

prioritize insights, followed by the selection of final recommendations. Often, policy and strategy research is conducted before or separately from the team creating and building a product. Policy and strategy research conducted in isolation may not provide practical and usable recommendations that can be taken forward during product development.

Limitations

We identified 3 key limitations of this project. First, no literature review was conducted to inform the research. Owing to the novelty of creating medical imaging platforms for AI development, we instead decided to conduct a scan of potential international efforts via targeted Google, Google Scholar, and social media searches.

Second, the ICR reliability indicates some variation in coding assignments between the 2 researchers (KK and EP). Coding variability could be attributed to (1) the level of experience analyzing qualitative research and (2) the depth of understanding of the topics discussed by the interview participants. It is important to note that the resultant ICR of 0.41 indicates moderate reliability, which falls within tolerance as outlined by Landis and Koch [49] and O'Connor and Joffe [12].

Third, the study did not delve into the role and importance of postmarket monitoring or surveillance. In some interviews, it appears that this topic was not top of mind as they were working on initiatives that were in the beginning stages and algorithms were not yet actively deployed into the market for clinical use. However, since the completion of this project, the NHS AI Lab has funded the United Kingdom Medicines and Healthcare products Regulatory Agency to deliver several work packages, including updating legislation to require more robust postmarket surveillance for software as a medical device [54].

Conclusions

Policy makers and digital developers internationally are chasing the potential for AI and ML algorithms to transform health care, with medical imaging seen as low-hanging fruit for realizing this ambition. Algorithms in health care are not confined to national borders, so how this ambition is realized by each country is particularly important. This paper outlines work undertaken by the NHS AI Lab to ensure that the investment in and creation of a generalized national medical imaging platform for the innovation and deployment of AI and ML algorithms in England is informed by international experience.

Acknowledgments

First, the authors would like to thank the stakeholders from each initiative for participating in this research. The authors learned a lot from each and every one of them and value their contributions. Second, the authors would like to thank the NHS AI Lab at NHS England, formerly at NHSX, for supporting the publication of this policy research and embedding the recommendations into the decision-making process for England's national imaging platform efforts. Finally, the authors would like to acknowledge Abhishek Mishra, who supported the earlier stages of the research while in a PhD intern placement at the NHS AI Lab and was funded by a Wellcome Trust doctoral scholarship. All research was conducted by staff members employed by or deployed to NHSX. No external funding was received to conduct the research. DC, Director of AI at the NHS AI Lab, NHS England, is the guarantor of the publication.

Authors' Contributions

KK conceptualized and supervised all stages of this project, including securing project resources, data curation, and project administration. DC was the main NHSX stakeholder and lead for the conceptualization and development of the National COVID-19 Chest Imaging Database and national artificial intelligence imaging platform. KK developed the research methodology with input from Abhishek Mishra and conducted this research alongside Abhishek Mishra and EP. EP and KK developed the discussion guide and deductive thematic analysis coding framework for the semistructured interviews. KK was the lead interviewer, and EP was the second interviewer and notetaker. KK and EP developed the workshop materials. KK was the lead workshop facilitator with support from EP. Transcription and translation services were provided by Prestige Network. KK and EP completed the thematic analysis and data synthesis. KK wrote the first draft of the manuscript. All the authors contributed to the drafting and editing of the manuscript and have approved the final version.

Conflicts of Interest

KK and EP were working at NHSX at the time of the study. DC was employed at NHSX at the time of the study and at NHS England at the time of writing.

Multimedia Appendix 1

Template discussion guide.

[[DOCX File, 19 KB - ai_v3i1e51168_app1.docx](#)]

Multimedia Appendix 2

Deductive thematic and coding framework.

[[DOCX File, 34 KB - ai_v3i1e51168_app2.docx](#)]

Multimedia Appendix 3

Description of international initiatives.

[\[DOCX File , 42 KB - ai_v3i1e51168_app3.docx \]](#)

Multimedia Appendix 4

Thematic synthesis.

[\[DOCX File , 207 KB - ai_v3i1e51168_app4.docx \]](#)

References

1. AICan 2020 CIFAR Pan-Canadian AI strategy impact report. Canadian Institute for Advanced Research. 2020. URL: <https://cifar.ca/wp-content/uploads/2020/11/AICan-2020-CIFAR-Pan-Canadian-AI-Strategy-Impact-Report.pdf> [accessed 2020-09-10]
2. Australia's AI action plan. Commonwealth of Australia. 2021 Jun. URL: https://wp.oecd.ai/app/uploads/2021/12/Australia_AI_Action_Plan_2021.pdf [accessed 2020-09-10]
3. National strategy for artificial intelligence. National Institution for Transforming India Aayog. 2018. URL: <https://niti.gov.in/sites/default/files/2019-01/NationalStrategy-for-AI-Discussion-Paper.pdf> [accessed 2020-09-10]
4. Data saves lives: reshaping health and social care with data. Department of Health and Social Care, Government of UK. 2022 Jun 15. URL: <https://www.gov.uk/government/publications/data-saves-lives-reshaping-health-and-social-care-with-data/data-saves-lives-reshaping-health-and-social-care-with-data> [accessed 2020-09-10]
5. HHS Artificial Intelligence (AI) strategy. US Department of Health and Human Services. 2022 Jan 10. URL: <https://www.hhs.gov/about/agencies/asa/ocio/ai/strategy/index.htm> [accessed 2020-09-20]
6. National COVID-19 Chest Image Database (NCCID). NHSX & GitHub. URL: <https://nhsx.github.io/covid-chest-imaging-database/> [accessed 2020-09-01]
7. Cushman D, Berka R, Bertolli O, Williams P, Schofield D, Joshi I, et al. Towards nationally curated data archives for clinical radiology image analysis at scale: Learnings from national data collection in response to a pandemic. Digit Health 2021;7:20552076211048654 [FREE Full text] [doi: [10.1177/20552076211048654](https://doi.org/10.1177/20552076211048654)] [Medline: [34868617](https://pubmed.ncbi.nlm.nih.gov/34868617/)]
8. The multi-agency advisory service (MAAS) - AI regulation - NHS transformation directorate. National Health Service, UK. URL: <https://transform.england.nhs.uk/ai-lab/ai-lab-programmes/regulating-the-ai-ecosystem/the-multi-agency-advice-service-maas/#about> [accessed 2020-10-01]
9. Groves L. Algorithmic impact assessment: a case study in healthcare. Ada Lovelace Institute. 2022 Feb 8. URL: <https://www.adalovelaceinstitute.org/report/algorithmic-impact-assessment-case-study-healthcare/> [accessed 2022-04-30]
10. Joshi I, Morley J. Artificial Intelligence: how to get it right: putting policy into practice for safe data-driven innovation in health and care. National Health Service X. 2019 Jan 01. URL: <https://transform.england.nhs.uk/ai-lab/explore-all-resources/understand-ai/artificial-intelligence-how-get-it-right/> [accessed 2023-11-30]
11. Aguilar FJ. Scanning the Business Environment. New York, NY: MacMillan Co; 1967.
12. O'Connor C, Joffe H. Inter-coder reliability in qualitative research: debates and practical guidelines. Int J Qual Methods 2020 Jan 22;19:160940691989922 [FREE Full text] [doi: [10.1177/160940691989922](https://doi.org/10.1177/160940691989922)]
13. How the discovery phase works. Government Digital Service, UK. 2021 Jun. URL: <https://www.gov.uk/service-manual/agile-delivery/how-the-discovery-phase-works> [accessed 2020-11-30]
14. Nominal Group Technique (NGT) - nominal brainstorming steps. American Society for Quality. 2020. URL: <https://asq.org/quality-resources/nominal-group-technique> [accessed 2020-10-30]
15. Anderson N, McKhann E. How to write compelling user research insights in 6 steps. Dscout. 2020. URL: <https://dscout.com/people-nerds/writing-user-insights> [accessed 2021-03-10]
16. Ip S, Liu T, Hodgett S. Machine learning and big data laws and regulations. Global Legal Insights. 2021. URL: <https://www.globallegalinsights.com/practice-areas/ai-machine-learning-and-big-data-laws-and-regulations/canada> [accessed 2020-10-01]
17. Innovation, science and economic development canada programs strategic innovation fund. Innovation, Science and Economic Development Canada, Government of Canada. 2022 Dec. URL: <https://ised-isde.canada.ca/site/strategic-innovation-fund/en> [accessed 2020-09-03]
18. Webster G. Full translation: China's 'new generation artificial intelligence development plan' (2017). New America. 2017 Aug 01. URL: <https://www.newamerica.org/cybersecurity-initiative/digichina/blog/full-translation-chinas-new-generation-artificial-intelligence-development-plan-2017/> [accessed 2020-09-30]
19. Feng C. China enhances smart health care with first national medical image database. South China Morning Post. 2020. URL: <https://www.scmp.com/tech/policy/article/3102534/china-enhances-smart-health-care-first-national-medical-image-database> [accessed 2020-11-30]
20. Sindermann C, Sha P, Zhou M, Wernicke J, Schmitt HS, Li M, et al. Assessing the attitude towards artificial intelligence: introduction of a short measure in German, Chinese, and English language. Künstl Intell 2020 Sep 23;35(1):109-118 [FREE Full text] [doi: [10.1007/s13218-020-00689-0](https://doi.org/10.1007/s13218-020-00689-0)]

21. Handley L. Chinese people are the most optimistic about the impact of AI on jobs. CNBC. 2018 Feb. URL: <https://www.cnbc.com/2018/02/07/chinese-people-are-the-most-optimistic-about-the-impact-of-ai-on-jobs.html> [accessed 2020-10-02]
22. Meinhardt C. The hidden challenges of China's booming medical AI market. China Business Review. 2019 Jun. URL: <https://www.chinabusinessreview.com/the-hidden-challenges-of-chinas-booming-medical-ai-market-2/> [accessed 2022-12-02]
23. Meng Q, Mills A, Wang L, Han Q. What can we learn from China's health system reform? BMJ 2019 Jun 19;365:l2349 [FREE Full text] [doi: [10.1136/bmj.l2349](https://doi.org/10.1136/bmj.l2349)] [Medline: [31217222](https://pubmed.ncbi.nlm.nih.gov/31217222/)]
24. Basu M. Exclusive: Hong Kong's vision for artificial intelligence. GovInsider. 2017 Oct. URL: <https://govinsider.asia/intl-en/article/exclusive-hong-kongs-vision-for-artificial-intelligence> [accessed 2020-09-15]
25. AI, machine learning and big data and regulations 2020 Hong Kong. Global Legal Insights. 2020. URL: <https://www.globallegalinsights.com/practice-areas/ai-machine-learning-and-big-data-laws-and-regulations/hong-kong> [accessed 2020-09-07]
26. Ng E. AXA boosts technology spending in Hong Kong as health revenues power growth. South China Morning Post. 2019 Nov. URL: <https://www.scmp.com/business/companies/article/3038159/axa-boosts-spending-ai-data-analytics-hong-kong-health-revenues> [accessed 2023-11-01]
27. Moltu C, Stefansen J, Svisdahl M, Veseth M. [Withdrawn] Doing business in Hong Kong: Hong Kong trade and export guide. Department for International Trade, Government of UK. 2015. URL: <https://www.gov.uk/government/publications/exporting-to-hong-kong/exporting-to-hong-kong> [accessed 2020-09-01]
28. Mori P. Is digital health finally taking off in Japan. Intralink. 2019 Apr. URL: <https://www.intralinkgroup.com/en-GB/News/Blog/April-2019/Is-digital-health-finally-taking-off-in-Japan> [accessed 2020-11-11]
29. Society 5.0. Cabinet Office, Government of Japan. 2020. URL: https://www8.cao.go.jp/cstp/english/society5_0/index.html [accessed 2020-10-17]
30. Gagan O. Society 5.0: is infrastructure key to Japan's success? Raconteur. 2020 Mar. URL: <https://www.raconteur.net/global-business/society-5-0-infrastructure/> [accessed 2020-09-03]
31. Japan: forecast of digital healthcare market size 2026 by segment. Statista. 2020. URL: <https://www.statista.com/statistics/1030901/japan-digital-health-market-size/> [accessed 2020-09-16]
32. Ravisconi M. The medtech opportunity for Japanese companies. McKinsey. 2017 Nov. URL: <https://www.mckinsey.com/industries/life-sciences/our-insights/the-medtech-opportunity-for-japanese-companies> [accessed 2020-09-27]
33. National artificial intelligence strategy: advancing our smart nation journey. Smart Nation Digital Government Office, Singapore. 2019. URL: <https://www.smartnation.gov.sg/files/publications/national-ai-strategy.pdf> [accessed 2020-09-07]
34. National approach to artificial intelligence. Government Offices of Sweden. 2018. URL: https://wp.oecd.ai/app/uploads/2021/12/Sweden_National_Approach_to_Artificial_Intelligence_2018.pdf [accessed 2020-09-10]
35. Vision for eHealth 2025. Ministry of Health and Social Affairs, and Swedish Association of Local Authorities and Regions. URL: https://ehalsa2025.se/wp-content/uploads/2021/02/Strategy-2020-2022_eng.pdf [accessed 2020-09-08]
36. Data protected Sweden. Linklaters. 2022 Jun. URL: <https://www.linklaters.com/en/insights/data-protected/data-protected---sweden> [accessed 2020-09-08]
37. Tang H. The European landscape - Sweden. AI-Med. 2020 Mar. URL: <https://ai-med.io/features/the-european-landscape-sweden/> [accessed 2020-09-11]
38. Bilboe C. Healthtech startups in Sweden and the UK with the fastest growth. Sifted. 2020 Sep. URL: <https://sifted.eu/articles/healthtech-growth-sweden-uk/> [accessed 2023-10-02]
39. Vestin E. Machine learning and big data laws and regulations. Global Legal Insights. 2020. URL: <https://www.globallegalinsights.com/practice-areas/ai-machine-learning-and-big-data-laws-and-regulations/sweden> [accessed 2020-09-10]
40. Lessons from application of AI to 6 year patient data from a Swedish primary care center. Strikersoft. 2020. URL: <https://strikersoft.com/en/News/what-can-ai-do-for-primary-care-lecture-from-vitalis/> [accessed 2020-09-09]
41. Artificial intelligence for the American people. Trump White House Archive. 2020. URL: <https://trumpwhitehouse.archives.gov/ai/> [accessed 2020-11-30]
42. Reardon S. Rise of robot radiologists. Scientific American. 2020 Feb. URL: <https://www.scientificamerican.com/article/rise-of-robot-radiologists/> [accessed 2020-09-11]
43. Caldwell A. The University of Chicago is awarded a major federal contract to host a new COVID-19 medical imaging resource center. UChicago Medicine. 2020 Aug. URL: <https://www.uchicagomedicine.org/forefront/coronavirus-disease-covid-19/the-university-of-chicago-is-awarded-a-major-federal-contract-to-host-a-new-covid-19-medical-imaging-resource-center> [accessed 2020-09-06]
44. The North America artificial intelligence in healthcare. GlobeNewswire. 2020 Sep. URL: <https://www.globenewswire.com/news-release/2020/10/01/2101805/0/en/The-North-America-artificial-intelligence-in-healthcare-diagnosis-market-is-projected-to-reach-from-US-1-716-42-million-in-2019-to-US-32-009-61-million-by-2027.html> [accessed 2020-09-30]
45. The Digital Health and Discovery Platform (DHDP). Digital Health and Discovery Platform. 2021. URL: <https://www.dhdp.ca/> [accessed 2020-09-08]

46. Hospital authority data sharing portal. Hospital Authority & Data Collaboration Lab. 2020. URL: <https://www3.ha.org.hk/data/DCL/Index/> [accessed 2020-09-08]
47. Karpathakis K, Murphy L, Mishra A, Joshi I. AI for healthcare: creating an international approach together. Global Digital Health Partnership. 2020. URL: <https://gdhp.health/work-streams/policy-environments/#whitepapers> [accessed 2020-09-11]
48. Home page. The Medical Imaging Data Resource Center (MIDRC). 2020. URL: <https://www.midrc.org/> [accessed 2023-10-02]
49. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics* 1977 Mar;33(1):159-174. [Medline: [843571](#)]
50. Morley J, Murphy L, Mishra A, Joshi I, Karpathakis K. Governing data and artificial intelligence for health care: developing an international understanding. *JMIR Form Res* 2022 Jan 31;6(1):e31623 [FREE Full text] [doi: [10.2196/31623](#)] [Medline: [35099403](#)]
51. Wilson J, Herron D, Nachev P, McNally N, Williams B, Rees G. The value of data: applying a public value model to the English national health service. *J Med Internet Res* 2020 Mar 27;22(3):e15816 [FREE Full text] [doi: [10.2196/15816](#)] [Medline: [32217501](#)]
52. Blandford A, Gibbs J, Newhouse N, Perski O, Singh A, Murray E. Seven lessons for interdisciplinary research on interactive digital health interventions. *Digit Health* 2018 Feb;4:2055207618770325 [FREE Full text] [doi: [10.1177/2055207618770325](#)] [Medline: [29942629](#)]
53. Bloomfield PS, Clutton-Brock P, Pencheon E, Magnusson J, Karpathakis K. Artificial intelligence in the NHS: climate and emissions ☆, ☆ ☆. *J Clim Chang Health* 2021 Oct;4:100056. [doi: [10.1016/j.joclim.2021.100056](#)]
54. Software and AI as a medical device change programme - roadmap. Medicines & Healthcare products Regulatory Agency. 2023 Jun 14. URL: <https://www.gov.uk/government/publications/software-and-ai-as-a-medical-device-change-programme/software-and-ai-as-a-medical-device-change-programme-roadmap> [accessed 2020-09-12]

Abbreviations

AI: artificial intelligence

AIDA: Analytic Imaging Diagnostics Arena

DA: Devolved Administration

DHDP: Digital Health and Discovery Platform

GDHP: Global Digital Health Partnership

HADCL: Hospital Authority Data Collaboration Laboratory

ICR: intercoder reliability

MIDRC: Medical Imaging and Data Resource Center

ML: machine learning

NCCID: National COVID-19 Chest Imaging Database

NHS: National Health Service

PESTLE: Political, Economic, Social, Technological, Legal, and Environmental

TRE: trusted research environment

Edited by Y Huo; submitted 23.07.23; peer-reviewed by M Halling-Brown, Z Li; comments to author 15.08.23; revised version received 01.09.23; accepted 03.11.23; published 04.01.24.

Please cite as:

Karpathakis K, Pencheon E, Cushnan D

Learning From International Comparators of National Medical Imaging Initiatives for AI Development: Multiphase Qualitative Study
JMIR AI 2024;3:e51168

URL: <https://ai.jmir.org/2024/1/e51168>

doi: [10.2196/51168](#)

PMID:

©Kassandra Karpathakis, Emma Pencheon, Dominic Cushnan. Originally published in JMIR AI (<https://ai.jmir.org>), 04.01.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR AI, is properly cited. The complete bibliographic information, a link to the original publication on <https://www.ai.jmir.org/>, as well as this copyright and license information must be included.

Original Paper

Framework for Ranking Machine Learning Predictions of Limited, Multimodal, and Longitudinal Behavioral Passive Sensing Data: Combining User-Agnostic and Personalized Modeling

Tahsin Mullick¹, MEng; Sam Shaaban², MBA; Ana Radovic³, MD, MSc; Afsaneh Doryab¹, PhD

¹Department of Systems and Information Engineering, University of Virginia, Charlottesville, VA, United States

²NuReIm, Pittsburgh, PA, United States

³Department of Pediatrics, University of Pittsburgh, Pittsburgh, PA, United States

Corresponding Author:

Tahsin Mullick, MEng

Department of Systems and Information Engineering

University of Virginia

Olsson Hall, 151 Engineer's Way

Charlottesville, VA, 22903

United States

Phone: 1 4349245393

Email: tum7q@virginia.edu

Abstract

Background: Passive mobile sensing provides opportunities for measuring and monitoring health status in the wild and outside of clinics. However, longitudinal, multimodal mobile sensor data can be small, noisy, and incomplete. This makes processing, modeling, and prediction of these data challenging. The small size of the data set restricts it from being modeled using complex deep learning networks. The current state of the art (SOTA) tackles small sensor data sets following a singular modeling paradigm based on traditional machine learning (ML) algorithms. These opt for either a user-agnostic modeling approach, making the model susceptible to a larger degree of noise, or a personalized approach, where training on individual data alludes to a more limited data set, giving rise to overfitting, therefore, ultimately, having to seek a trade-off by choosing 1 of the 2 modeling approaches to reach predictions.

Objective: The objective of this study was to filter, rank, and output the best predictions for small, multimodal, longitudinal sensor data using a framework that is designed to tackle data sets that are limited in size (particularly targeting health studies that use passive multimodal sensors) and that combines both user agnostic and personalized approaches, along with a combination of ranking strategies to filter predictions.

Methods: In this paper, we introduced a novel ranking framework for longitudinal multimodal sensors (FLMS) to address challenges encountered in health studies involving passive multimodal sensors. Using the FLMS, we (1) built a tensor-based aggregation and ranking strategy for final interpretation, (2) processed various combinations of sensor fusions, and (3) balanced user-agnostic and personalized modeling approaches with appropriate cross-validation strategies. The performance of the FLMS was validated with the help of a real data set of adolescents diagnosed with major depressive disorder for the prediction of change in depression in the adolescent participants.

Results: Predictions output by the proposed FLMS achieved a 7% increase in accuracy and a 13% increase in recall for the real data set. Experiments with existing SOTA ML algorithms showed an 11% increase in accuracy for the depression data set and how overfitting and sparsity were handled.

Conclusions: The FLMS aims to fill the gap that currently exists when modeling passive sensor data with a small number of data points. It achieves this through leveraging both user-agnostic and personalized modeling techniques in tandem with an effective ranking strategy to filter predictions.

(JMIR AI 2024;3:e47805) doi:[10.2196/47805](https://doi.org/10.2196/47805)

KEYWORDS

machine learning; AI; artificial intelligence; passive sensing; ranking framework; small health data set; ranking; algorithm; algorithms; sensor; multimodal; predict; prediction; agnostic; framework; validation; data set

Introduction

Background

Mobile and wearable sensing has garnered increasing interest in areas of physical health [1,2], mental health [3-5], and activity recognition [6,7]. Multimodal passive sensing accommodates data collection without disrupting the human routine, allowing it to be an important tool to understand human behavior. However, passive sensing, unlike other forms of data, encounters common fundamental challenges in mobile health studies pertaining to physical and mental health. These challenges include small data sets, noisy or sparse data, and sensor selection criteria. Next, we explain these challenges and discuss how our framework can help in alleviating them.

One of the primary challenges in passive sensing studies is small data sets. These arise due to limitations in the sample size of participants, the study duration, and ground truth restrictions. In this study, we explored this challenge from the viewpoint of studies conducted on passive sensing. Studies related to physical health (eg, [1,2]) have investigated dietary behavior with the help of passive sensing. Participant sample sizes in Rabbi et al [1,2] were 17 and 16, respectively, which is a limited participant count. This type of data limitation is even more prominent in mental health research that relies on passive sensing. Studies on depression [3] and schizophrenia [4], for example, had participant sample sizes of 28 and 5, respectively. The limited data sets in passive sensing research are also a factor of the study duration. To understand this, we can observe the duration of study. For example, the study duration in Rabbi et al [1,2] was 21 and 98 days, respectively, while the study by Canzian and Musolesi [3] lasted for 70 days and that by Difrancesco et al [4] was limited to only 5 days. The limitation in data led researchers away from using complex deep learning (DL) models, as demonstrated in previous studies [1-4]. This is because DL models have more hyperparameters and succumb to overfitting due to memorization of the data the models are trained on [8]. In this study, we took inspiration from the existing work and selected specific traditional machine learning (ML) algorithms that are less susceptible to overfitting in small-data scenarios. However, unlike previous studies [1-4,9-17], we also ensured that our predictions were ranked based on 2 different modeling paradigms that further helped circumvent overfitting and also assisted in noise removal, as explained later.

The second challenge commonly faced when tackling passive sensor data is that of sparsity or noise. This challenge arises due to signal inconsistencies and noise in sensor data collection because of software issues, data sync, or hardware problems. Discussions of sparsity and the negative effect it has on modeling have been previously documented [7,18-20]. These studies have presented an overview of the passive sensing landscape and highlighted the role signal inconsistencies can play in predictive modeling of passively sensed data. The fact

that data are noisy, especially in the case of wearable sensors, was mentioned by Plötz [18]. Cornet and Holden [19] reported that a lack of sensor precision leads to sparsity, and Xu et al [20] documented the level of noise in data that prevents user-agnostic models from generalizing well. Our proposed framework attempts to reduce the effect of noise by forming a balance between predictions from user-agnostic modeling paradigms and personalized modeling paradigms. In addition, choosing specific ML algorithms, such as Extreme Gradient Boosting (XGBoost), Adaptive Boosting (AdaBoost), elastic-net, and extra-tree, and ranking predictions from them help lessen the impact of sparsity [21-24].

Sensor selection is the third type of challenge that has not received significant attention in passive or mobile sensing literature. Studies have tested various feature combinations mainly in the light of performing feature selection or feature reduction [25]. Joshi and Boyd [26] and Altenbach et al [27], for example, used heuristic-based convex optimization to select sensors from an array of sensors. However, both these studies were purely from the perspective of sensor placement. They did not investigate which combination of sensors provided the best outcome for prediction-based modeling and were more in favor of wireless sensor network establishment. Mobile or wearable devices are laced with multiple sensors, and building and knowing which sensors create optimum models are vital particularly to mental and physical health-related studies. Through our framework, we present a way to test combinations of sensor data and derive and rank predictions from among those combinations, allowing investigators to understand which combinations of sensor data yield the best predictions for their passive sensing experimental setup.

All the aforementioned challenges are common to passive sensing data sets. However, they exhibit significant presence in mental and physical health-related studies [3,4]. Xu et al [20] talked of the general sequence of steps researchers take to build models and the struggles of working with passively sensed data. A strong framework to yield the best predictions can prove to be beneficial to the community at large and bring about greater insight from studies conducted with small data sets.

In this paper, we present our ML modeling and ranking framework to address these challenges. The framework is designed to induce improved predictions for multimodal sensing. It balances both user-agnostic and personalized modeling of small data sets encountered often in mental and physical health-based studies. Our framework makes the following contributions: (1) prediction filtering and ranking through tensor-based aggregation of small, multimodal sensing data sets, (2) sensor combination selection to derive the best predictions, and (3) a reduction in overfitting predictions due to limited data and noise through ensembling of user-agnostic and personalized modeling strategies.

Importantly, it should be noted that by the size of the data set, we refer to the final data sets where raw sensor readings are

aggregated into intervals to align with the sampling frequency of ground truth data. In this work, we defined small data sets as those comprising fewer than 1000 data points for training ML models. Sparse or noisy data sets were those that either consisted of many zero entries or data sets for which highly varying sensor values were observed among different participants in the study.

We evaluated the framework through its performance in the context of predicting changes in depression severity in a group of adolescent patients. The results showed the framework's ability to use multiple modeling approaches for providing robust predictions in critical cases, such as mental health.

Passive sensing data for human behavior modeling are different from other data formats, such as images, audio, or normal tabular data. Researchers in the field of passive sensing agree that passive sensing data have some common properties, such as they are time series data, multimodal, longitudinal, nonlinear, and noisy, as previously discussed [20]. Xu et al [20] also emphasized the researcher's need for tools that can help ease the time lost in traversing the common pitfalls of passively sensed data. Our work endeavors to resolve such pitfalls for cases where passive sensing data are limited. Next, we discuss the related work highlighting the state of the art (SOTA) in passively sensed small, multimodal data sets.

Related Work

Despite the growing body of work using multimodal passive sensing in physical and mental health applications [28-32], there exists scope for improvement in small-data scenarios.

In this section, we underline what exists in the current SOTA and why we need a ranking-based framework to address scenarios with small data sets. Keeping in line with our contribution, it will prove beneficial to present the current SOTA through understanding:

- How traditional ML algorithms are applied in the context of passive sensing
- Why complex DL models do not work well in limited data scenarios
- How ensemble modeling has been adapted in passive sensing studies
- What the role of data fusion is in modeling passive sensing data

Traditional Machine Learning Algorithms Applied in Passive Sensing

Traditional ML algorithms have been applied to passive sensing in the space of human activity recognition (HAR) [9-11], general health [12-15], and mental health [3,16,17]. A deeper dive into the studies reveals some common takeaways that include the following:

- All of them test multiple ML algorithms, followed by selecting predictions based on the overall chosen validation metric.
- They all follow a singular modeling strategy, resorting to either user-agnostic or personalized modeling.
- Cross-validation (CV) is either K-fold or leave-one-out CV.

This is a repetition of steps that authors in the field make independently and is discussed extensively in the highlighted literature presented in Table 1. Following a single modeling strategy is restricting as choosing to follow a user-agnostic approach exposes the model to a greater degree of noise due to the heterogeneity in sensor values among participants, while solely following a personalized approach reduces data availability further as the model learns from individuals' data rather than the general population data. Our endeavor through this ranking framework is to combine both the approaches, while using traditional ML algorithms.

Table 1. Summary of SOTA^a literature using traditional ML^b for passive sensing, with special focus on CV^c, the overall modeling strategy, and ML algorithms.

Study	Application	CV	Modeling strategy	ML algorithm
Kwapisz et al [9]	HAR ^d	10-fold	User agnostic	DT ^e , LR ^f , MLP ^g
Shukla et al [10]	HAR	5-fold	User agnostic	KNN ^h , SVM ⁱ
Chen and Chen [11]	HAR	10-fold	User agnostic	RF ^j , SVM, KNN
Huang et al [12]	Sleep	10-fold	User agnostic	SVM
Montanini et al [13]	Sleep	K-fold/leave 1 out	User agnostic/personalized	KNN, DT, RF, SVM
Teng et al [14]	Parkinson's tremors	5-fold	User agnostic	XGBoost ^k , DT, RF
Azam et al [15]	Breath	K-fold	User agnostic	SVM
Canzian and Musolesi [3]	Depression	Leave 1 out	User agnostic	SVM
Grunerbl et al [16]	Bipolar disorder	K-fold	User agnostic/personalized	NB ^l , KNN, DT
Saeb et al [17]	Depression/anxiety	10-fold	User agnostic	XGBoost, DT

^aSOTA: state of the art.

^bML: machine learning.

^cCV: cross-validation.

^dHAR: human activity recognition.

^eDT: decision tree.

^fLR: linear regression

^gMLP: multilayer perceptron

^hKNN: K-nearest neighbor

ⁱSVM: support vector machine.

^jRF: random forest

^kXGBoost: Extreme Gradient Boosting

^lNB: naive Bayes

Limitation of Deep Learning in Small-Data Scenarios

A common replacement for traditional ML algorithms is DL. Here, we explain why DL models are not ideal solutions for the problem addressed in this study. DL models have gained immense popularity in the literature [33]. Their power lies in modeling the nonlinearity and noisy nature of passively sensed data. DL has a toolkit of strategies to handle small data that includes data augmentation [1], transfer learning [19], and ensembling [29]. However, the size of a small data set in DL studies ranges from 1000 to 10,000 training points [18]. This is unlike the ranking framework presented in this paper, which has been designed for data sets with fewer than 1000 data points. Therefore, despite their superiority in modeling larger passive sensing data sets, the performance of DL models suffers in cases where study data are limited and in the hundreds. The complexity of DL models results in overfitting to small data sets [14]. In this paper, we worked to solve the problem of limiting data by providing researchers with a reproducible way to run multiple models and select the best predictions from among them. By using traditional ML in conjunction with ranked predictions from user-agnostic and personalized models, the issue of overfitting due to model complexity is dealt with in the proposed work.

Ensemble Learning to Build Robust Models for Passive Sensing Data

Among the different ways of dealing with overfitting, ensemble learning has been instrumental. Ensemble ML is a widely used approach in passive sensing studies [14,17,34,35]. It mainly exists in the form of boosting [6,14,17,34], bagging [14,16], weighted ensembles [35], and max voting [36] ML algorithms. Ensemble learning presents better results in terms of evaluation metrics. Ensemble learners are trained using a single modeling strategy. Therefore, they are either personalized ensembles [35], which allows learners to derive interesting artifacts at personal levels, or user-agnostic ensembles [14,17,34,36-38], which only generate macrolevel information. Our contribution through the ranking framework is to provide a balance of both macrolevel patterns and user-specific patterns through a weighted ensemble of both approaches. Ensembling in this manner will allow us to reduce the noise that is picked up due to varying sensor values among users and account for user-specific patterns through the predictions on personalized data.

Role of Data Fusion in Passive Sensing Studies

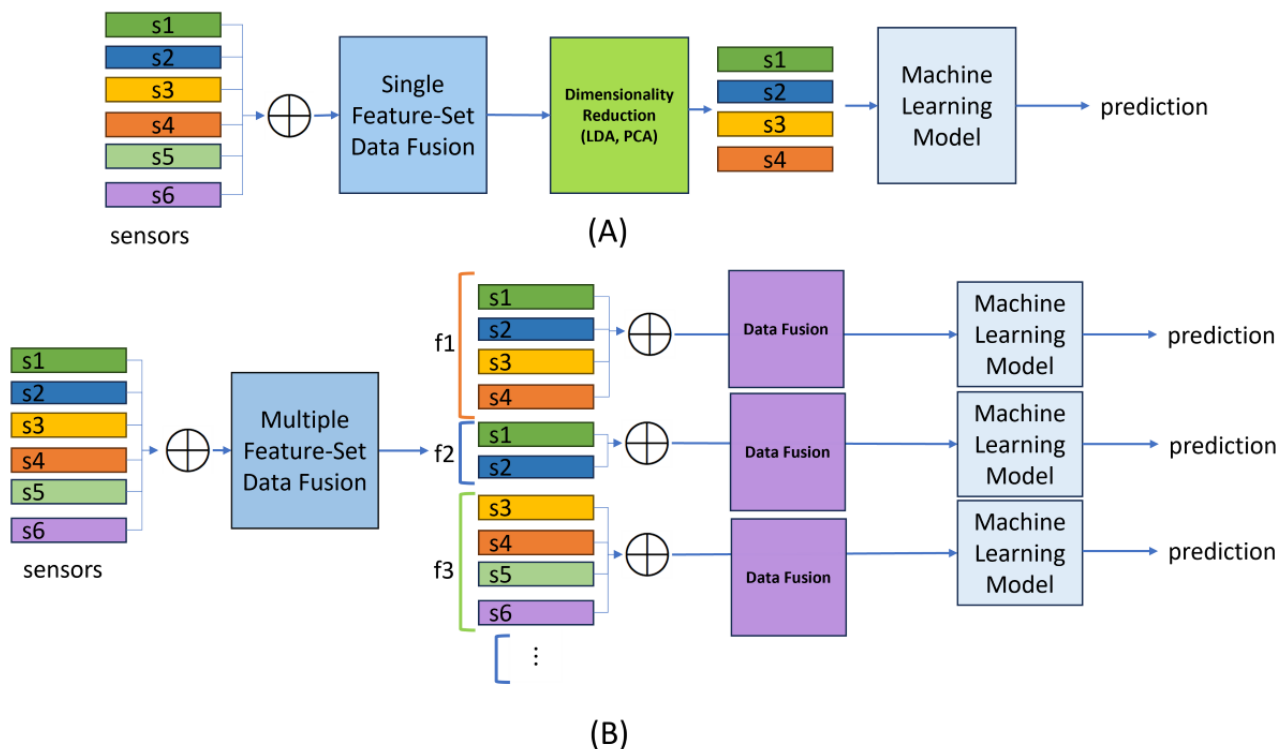
The use of data fusion in passive sensing has seen a steady growth due to the use of multimodal sensors in passive sensing studies. Earlier studies were often restricted to single sensors, which were then manipulated to obtain a handful of features. For example, Canzian and Musolesi [3] primarily used GPS sensor data, while Kwapisz et al [9] only opted for an

accelerometer to base their predictive modeling. The way data fusion is approached has a common link among the surveyed studies in the current literature. The studies have applied feature-level fusion [10,39-43], where fusion takes place after feature extraction from raw signals. A single feature set is generated and then passed on to dimensionality reduction, such as linear discriminant analysis (LDA) [10] or principal component analysis (PCA) [40-42]. The focus in these papers tends to be a reduction in dimension, without trying to study the impact of multiple distinct feature combinations. In comparison, our contribution of feature selection focuses on

studying the relationship between each group of sensors by creating multiple feature sets based on sensor availability. This will allow us to select the best set of features to work with for a specific type of study. An illustration of the difference in the existing literature and our feature fusion approach is shown in Figure 1 [10,39-43].

Overall, our ranking framework is motivated to aid researchers in situations in which data sets are small, sparse, or noisy and multimodal by taking advantage of its multiple model generation and the balanced outcome of the best predictions.

Figure 1. (A) Data fusion approach in the current literature and (B) proposed FLMS data fusion approach, where s1-s6 represent distinct sensors and f1-f3 represent feature set combinations, which were then fused prior to ML modeling. FLMS: framework for longitudinal multimodal sensors; LDA: linear discriminant analysis; ML: machine learning; PCA: principal component analysis.



Methods

Ethical Considerations

The data collection was approved by the Institutional Review Board of the University of Pittsburgh Human Research Protections Office (STUDY18120176).

Data Description

The study used passive sensing data and is presented through the lens of depression change prediction among adolescents. The data set comprised 55 adolescents from 12 to 17 years old, with an average age of 15.5 (SD 1.5) years. The AWARE app was used to collect the participants' smartphone and Fitbit data. The data completeness rate for AWARE and Fitbit was, on average, 65.11% and 30.36%, respectively. The levels of completeness echoed the difficulty in collecting passive sensing data. Smartphone and Fitbit data were collected from each participant over 24 weeks.

The 9-item Patient Health Questionnaire (PHQ-9) [44] was used to collect weekly self-reports of depression severity from the participants. The questionnaire consists of a set of 9 questions, which can be scored from 0 to 3, giving a score range of 0-27. We used PHQ-9 scores as the ground truth to compare the prediction accuracy of our models.

Relation of Sensor Data to Mental Health

Raw sensor data, including calls, location, conversation, screen usage, Wi-Fi, steps, sleep, and heart rate, were processed, and relevant features were extracted at daily intervals. We used RAPIDS [45] to extract 72 features from the sensors. The existing literature [3,46-51] shows how location [3,46,49,50,52], calls [48,53], screen usage [46,54,55], conversations [55-58], Wi-Fi [48,59], steps [60], and heart rate [61] can be effective in predicting mental health behavior. Studies [3,46,49,50] have used location sensors, such as the GPS, and shown a strong relation to depressive symptom severity. Clinical measures, such as the PHQ-9 [44], the PHQ-8 [62], the Hamilton Rating Scale for Depression (HAM-D) [63], and the Hamilton Rating

Scale for Anxiety (HAM-A) [64], have been used as target labels for prediction using sensor-based features, establishing a proof of association between sensor features and mental health predictions. Studies [47,48,51,54,60] have used multimodal sensors of smartphones that included the sensors we chose for this study: calls, location, conversation, screen usage, Wi-Fi, Fitbit steps, and Fitbit heart rate. In the *Results* section, we further elaborate on the feature engineering from each of the sensors. The validity of using the sensors to predict mental health, in particular the choice of sensors, was motivated by the aforementioned studies, which showed strong predictive capability of sensors in the area of mental health prediction.

Framework Design and Modeling

We proposed a framework for longitudinal multimodal sensors (FLMS) as a ranking framework to rigorously handle longitudinal, multimodal sensor data and incorporate different analysis and modeling strategies suited for small and sparse time series data sets to produce better results. The FLMS incorporates 4 stages to improve, rank, and filter data set predictions (see Figure 1):

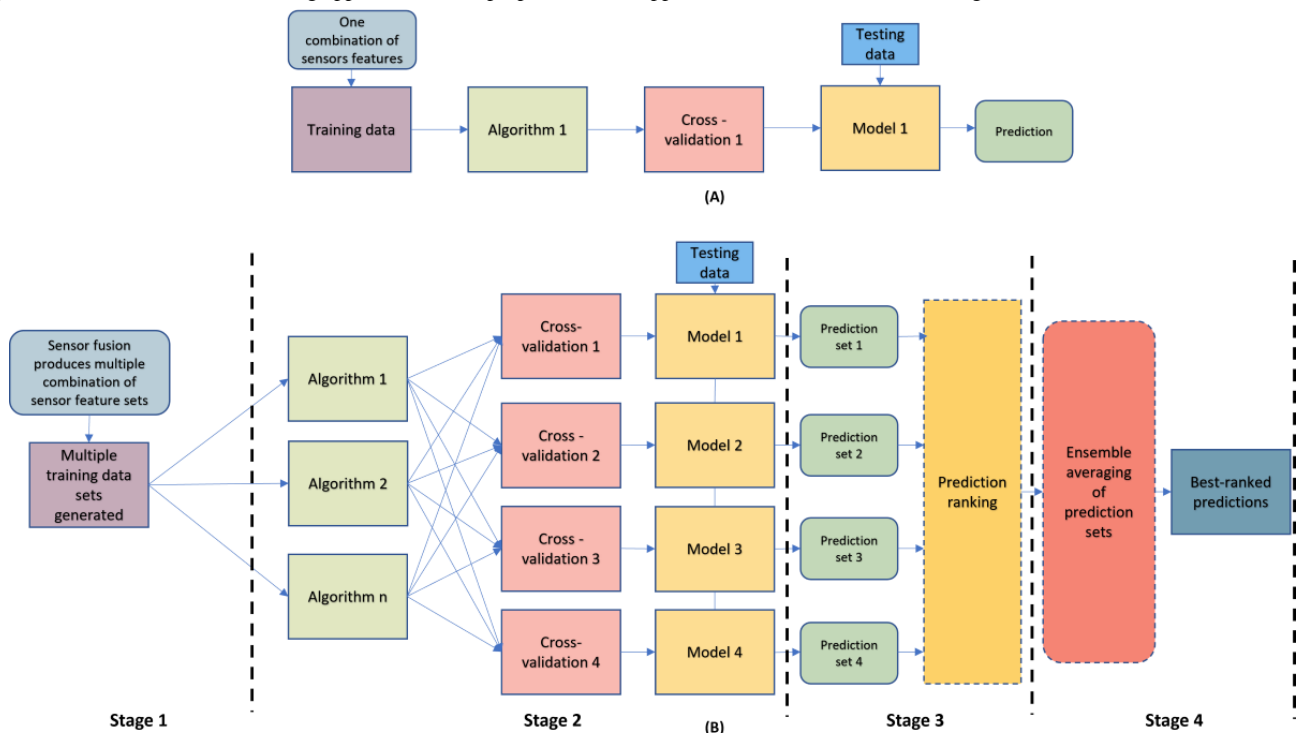
- Stage 1: multimodal sensor fusion to explore the data set from multiple views and to identify the minimum number of sensors necessary to yield a good prediction. It also addresses sparsity.
- Stage 2: ML modeling with combined user-agnostic and personalized approach. This stage is designed to leverage

- user-agnostic and personalized predictions. The ML algorithms used in this stage were chosen due to their superior prediction capability in small-data scenarios and their ability to tackle sparse data sets.
- Stage 3: tensor-based aggregation and ranking leverage predictions from all fused combinations and modeling strategies to calculate more robust predictions.
- Stage 4: final prediction informed by the ensemble weighted average of both user-agnostic and personalized predictions to reduce the effect of overfitting in small data sets. This stage uses weights calculated via hamming distances to prevent any modeling approach from dominating the predictions.

A high-level view in Figure 2 illustrates how the FLMS is different from conventional ML approaches. Observing Figure 2A, we understand that the conventional modeling strategy uses a single algorithm with either a user-agnostic CV, where all users are included in the training and test sets, or a personalized CV strategy, where a single user’s data are used to derive predictions. However, Figure 2B displays how the FLMS uses different combinations of sensors as input data, followed by multiple algorithms and a combination of user-agnostic and personalized modeling. The modeling stage is followed by a ranking of predictions and finally an ensemble of the predictions to yield the final output.

A detailed explanation of the stages of the FLMS and their utility is provided next.

Figure 2. (A) Conventional modeling approach and (B) proposed FLMS approach. FLMS: framework for longitudinal multimodal sensors.



Stage 1: Multimodal Sensor Fusion

Stage 1 was designed for the early fusion of sensors at a feature level. Sensor fusions followed a combinatorial approach using $\binom{Z}{x}$, where Z is the total number of modalities available and x

is the number of sensors to fuse. Our case study had 6-sensor modalities that generated a set of 63 separate data sets calculated as $\binom{6}{x}$.

Data set preprocessing steps involved normalization and log transforms. Imputations to fill missing feature observations

were also conducted. The framework allowed for implementation of the K-nearest neighbor (KNN) algorithm for imputation, which is also the first level of defense against sparsity. The generated data sets were in 2D tabular data format. The sensor data were aggregated according to the granularity of the ground truth. Our case study collected PHQ-9 scores as an accepted depression measure. The total score range of the 9 questions was 0-27. This was collected on a weekly basis, and thus, our daily data were aggregated in weekly intervals.

Stage 2: ML Modeling With a Combined User-Agnostic and Personalized Approach

Stage 2 focused on modeling and predictions based on the data sets generated in stage 1. All stage 1 data sets were run through the modeling suite, which encompasses a series of ML algorithms and CV strategies to help build user-agnostic and personalized models.

The ML suite includes case-specific linear and nonlinear algorithms. For our case study on adolescent depression, we followed a regression-based approach, and therefore, we selected algorithms such as linear regression (LR), elastic-net, random forest (RF), AdaBoost, extra-tree, gradient boosting, and XGBoost. The algorithms were chosen based on (1) their performance in the existing literature when working with small data and robustness to sparsity, and (2) tree-based models, which were specifically chosen to provide added tractability for researchers to inspect which features mainly contributed to the models’ predictive capability. The algorithms were used in each modeling strategy. The predictions of the ML algorithms for each time unit were stored in arrays for each participant and

later used to select the best model for each participant. The best model selection strategy chose the model with the minimum error (in the case of regression) or the maximum accuracy (in the case of classification) among all algorithms. For example, among l number of regression algorithms, the best model was chosen as follows:

$$\boxed{\times}$$

(1)

,where alg refers to the algorithm with the lowest absolute sum error and $\text{pred}_m(\text{alg}_t)$ is the prediction made by an algorithm l at unit time t. The array of prediction by the best model was retained for each respective participant.

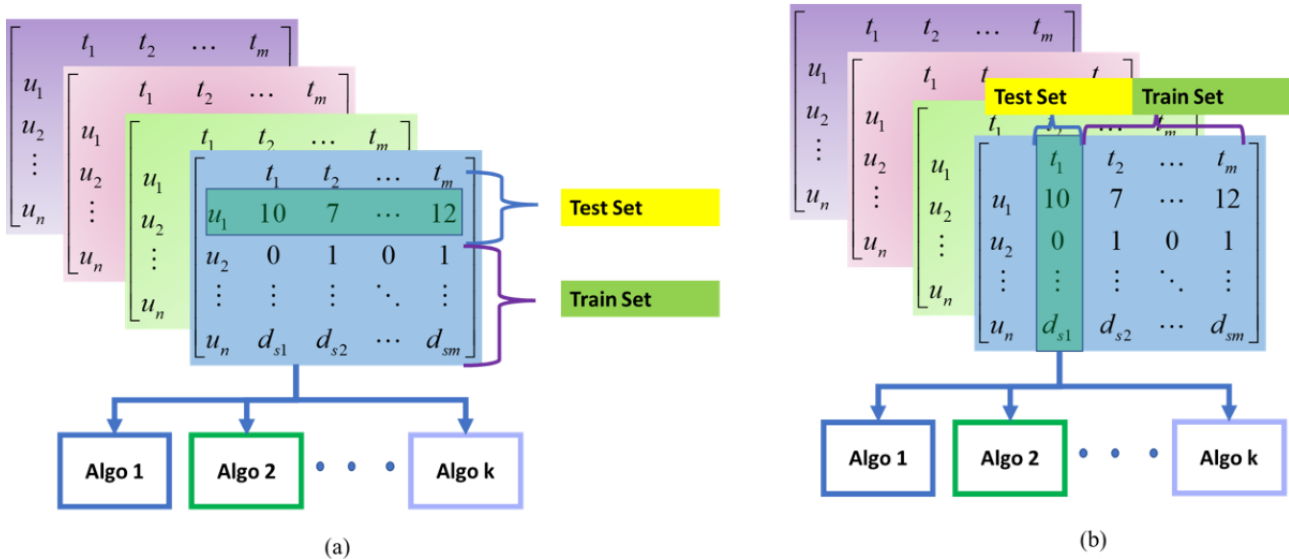
User-Agnostic Model Building

To leverage as much data as possible, we implemented the leave-one-participant-out (LOPO) and leave-time-unit-X-out (LTXO) strategies. This is illustrated in Figure 3A,B.

In LOPO, we held out all data from a single participant for validation and trained the model on other participants. This strategy reflected the cold start case where a new user started using the health app.

The LTXO is based on the unit of time for ground truth data (eg, a week). For training, we held out a given time unit of all participants and trained the model on the rest of the time units. This strategy evaluated the impact of time-specific segments of data on prediction. The training phase captures the similarity and variation of data during different time units to build user-agnostic models.

Figure 3. User-agnostic model building: (A) LOPO and (B) LTXO strategies. Algo: algorithm; LOPO: leave one participant out; LTXO: leave time unit X out.



Personalized Model Building

The personalized modeling strategy leverages each user’s historical and cross-time data samples in a sliding window and the leave-one-time-unit-out approach.

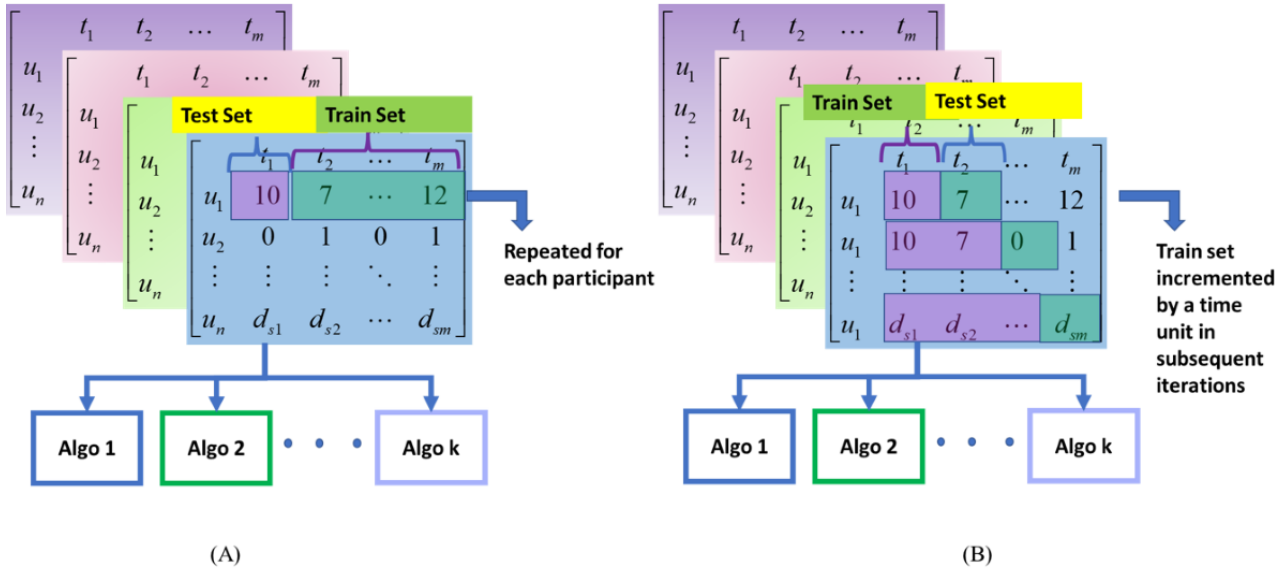
For each participant, the accumulated-time-unit (ATU) strategy built a model from X_t time units of data to predict X_{t+1} . For example, the model built from weeks 1 and 2 predicted depression in week 3. In the next iteration, the sliding window was increased by T time units (eg, 2 weeks) to repeat the model-building process. This process continued until the

maximum number of time units was reached. This method examined the forecasting capability of the framework.

The leave-one-time-unit-one-participant-out (LOTPO) strategy trained the models on all time units of a participant across time

to predict the target label for the current time unit. For example, for a participant with 10 weeks of data, we built a model from data in weeks 1-5 and weeks 7-10 to predict depression in week 6. This method evaluated the feasibility of past and future data for each participant to predict an outcome (Figure 4A,B).

Figure 4. Personalized model building: (A) LOTPO and (B) ATU strategies. Algo: algorithm; ATU: accumulated time unit; LOTPO: leave one time unit of participant out.



Stage 3: Tensor-Based Aggregation and Ranking

The output of stage 2 was a set of best prediction matrices for sensor fusion combinations, where each slot in the matrix represented prediction results for a participant in a particular time unit. We represented these predictions in the form of Z-dimensional tensors (Figure 5), where Z is the number of modalities being used. For example, a study with 6 modalities and 45 users over 24 weeks was represented in tensor form as (6, 45, 24). The tensor representation helped represent the high dimensionality of sensor combinations.

The predicted values for each slot across tensors were then aggregated using an aggregation function (eg, mean). This process took advantage of the stage 2 combinations to help reduce the error in prediction. For example, we aggregated predictions of 6 tensors (generated from 5-sensor fusion) into 1 tensor by calculating the mean of the predictions from the 6 combinations (see Figure 3). This was done for both user-agnostic and personalized models. The aggregated mean was calculated using the following equation:

$$\text{[Equation symbol]$$

(2)

,where M_{agg} is the aggregated mean, k is the total number of sensor combinations aggregated, i is the combination number, j is the corresponding time unit, and [Equation symbol] is the prediction across

each set of combinations. The data were now in a format where each 2D tensor represented a particular sensor fusion prediction set (Figure 6).

The predictions were next encoded into 0s and 1s to counter the large variance in the regression values from the original values. This logic can be set based on the type of ML problem the framework is being used to address. For example, in our case study, if the regressed change in depression score values was 0 or negative value, we classified it as 0, and if it was positive, we represented it as 1 (Figure 7).

The next step in this stage measured the hamming distance between the 0-1-encoded tensor and the true labels tensor, as shown in Figure 8. These hamming distances were then aggregated (D_u) for the respective 2D tensor as follows:

$$\text{[Equation symbol]$$

(3)

,where $d(p_i, a_i)$ is the hamming distance between unit time predictions p_i and the true value a_i . Based on the measured distance, we ranked and chose the best set of predictions. This metric helped inform the choice of weightage to associate with a particular modeling strategy. The hamming distance helped further reduce errors after encoding and filtered down to the best set of predictions from each strategy.

Figure 5. An example of tensor representation of 6-sensor fusion predictions.

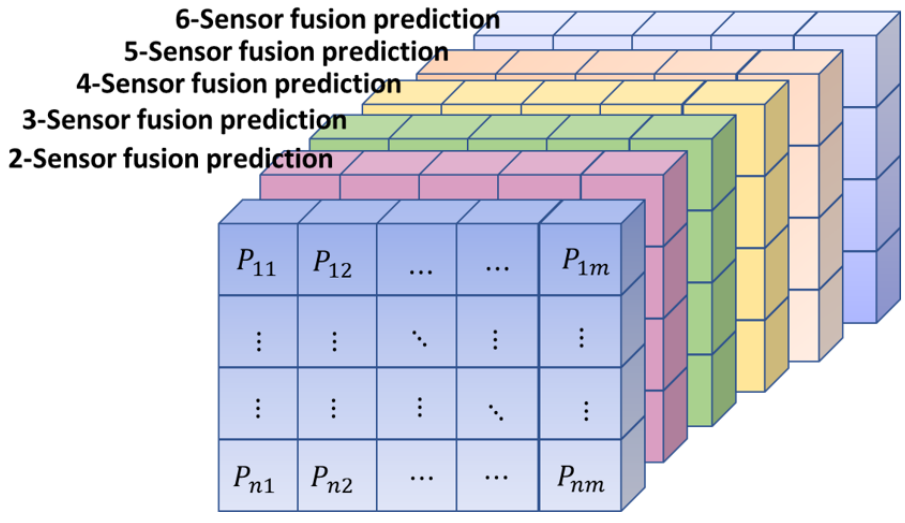


Figure 6. Instance of ATU where it shows how the mean aggregated prediction set is generated according to Equation (2). ATU: accumulated time unit; avg: average.

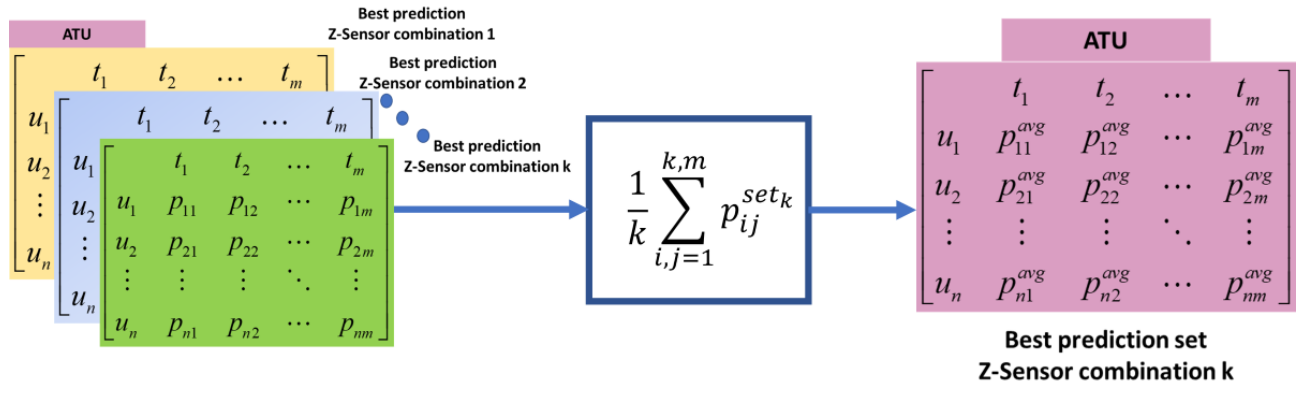


Figure 7. The 0-1 encoding process resolves dealing with large variances in regression values. ATU: accumulated time unit; LOPO: leave one participant out; LOTPO: leave one time unit of participant out; LTXO: leave time unit X out.

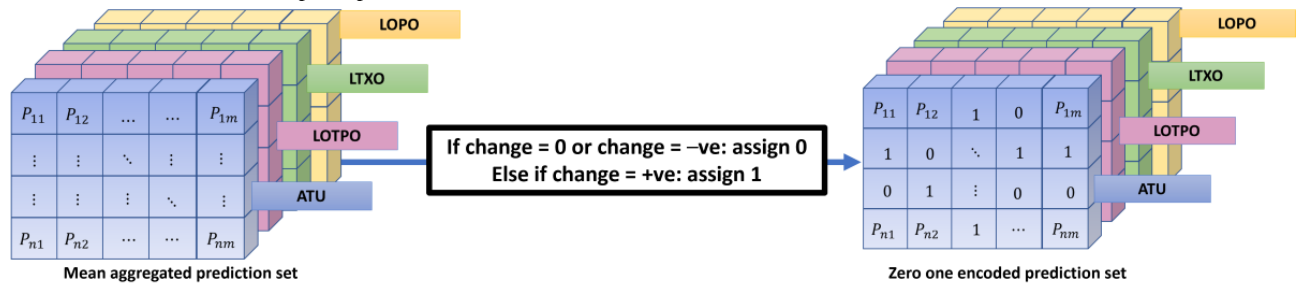
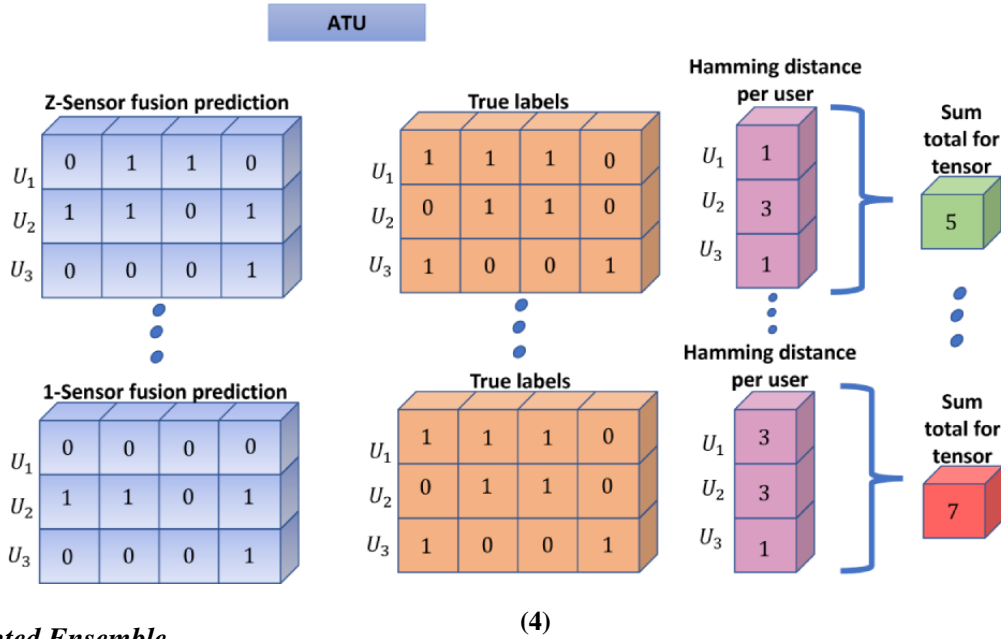


Figure 8. Hamming distance calculations reduce error and also determine the weight each of the 4 modeling approaches will contribute to stage 4's ensemble weighted average. ATU: accumulated time unit.



Stage 4: Weighted Ensemble

The final stage formed the most robust set of predictions via an ensemble weighted average approach, where weights were calculated based on the minimum hamming distances derived from each modeling strategy in stage 3 (Figure 9):

,where P_{ij} is the prediction tensor, w_k is the weight based on the minimum hamming distance, and i and j are the number of users and time units, respectively. The data were then encoded back to 0s and 1s. A complete version of the FLMS with all its stages is presented in Figure 10 (see Multimedia Appendix 1 for a higher quality image).



Figure 9. Ensemble average based on weights derived from the hamming distance to arrive at best-ranked predictions. ATU: accumulated time unit; LOPO: leave one participant out; LOTPO: leave one time unit of participant out; LTXO: leave time unit X out.

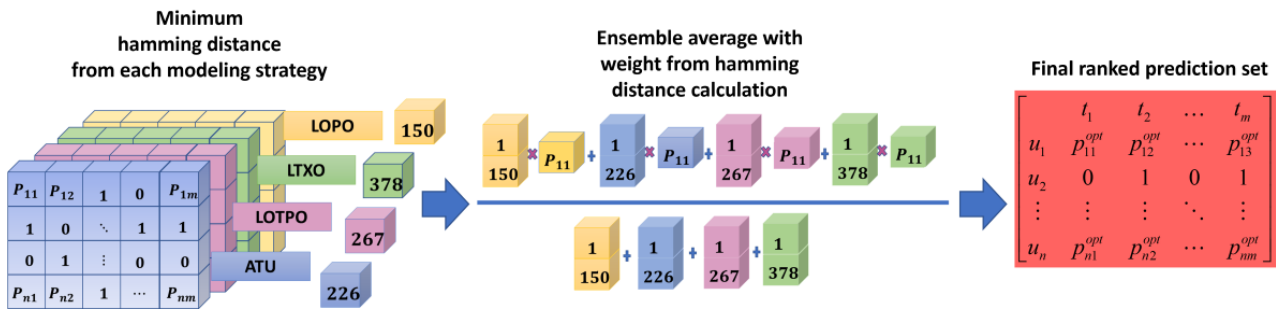
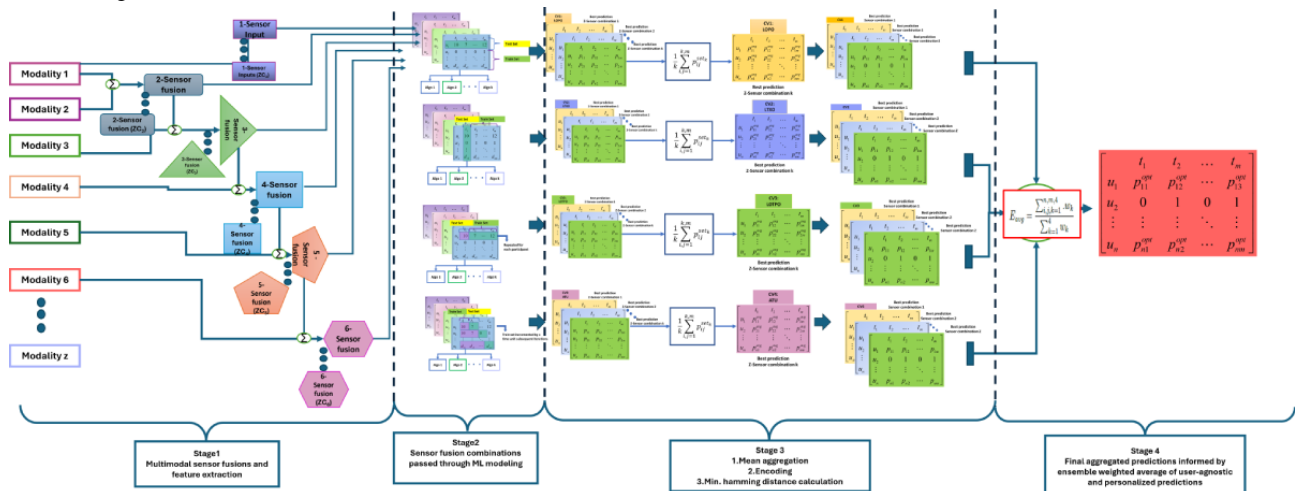


Figure 10. FLMS ranking overview. Algo: algorithm; ATU: accumulated time unit; avg: average; CV: cross-validation; FLMS: framework for longitudinal multimodal sensors; LOPO: leave one participant out; LOTPO: leave one time unit of participant out; LTXO: leave time unit X out; ML: machine learning.



Results

Stagewise Description of Framework Processing on an Adolescent Data Set

To evaluate the performance of the proposed FLMS, we used a depression data set of adolescents. This was a small data set, comprising noisy, multimodal sensor values from multiple participants—a suitable case study for our purpose of evaluating the performance of our proposed framework. Before presenting the experimental results, we first provide an understanding of how the adolescent data set was processed at each stage of the FLMS.

The passively sensed depression data set was longitudinal, with a varying number of observations per participant. The goal was to predict changes in the depression score. This was achieved by passing the small set of observations through our ranking framework, which processed, modeled, ranked, and output the best set of overall predictions based on multiple modeling approaches. A prediction of change in depression is difficult and becomes even more challenging when the amount of data provided to the ML algorithms is limited.

Stage 1 Outcome

As part of stage 1, daily data were aggregated in weekly intervals to align with weekly ground truth values. Based on our extensive exploratory data analysis (EDA), we set thresholds for sparsity and adopted KNN as the imputation strategy.

Our final data set consisted of 507 data points with 72 features, with an average of 13 weekly data points per participant. A series of data sets were then produced from an early fusion of 6-sensor features. Each data set retained 45 (81.8%) of the 55 participants. We had to drop 11 (20%) participants as they were missing more than 60% of their sensor data. The true depression state of the participants was given by the PHQ-9 weekly survey. The change in participant depression scores was calculated as $W_m - W_{m-1}$, where W_m is the score for the m-th week; this served as the ground truth for our analysis.

Stage 2: ML Modeling Outcome

The ML algorithms in stage 2 regressed on the change in the depression score, with positive changes exhibiting a rise in the depression score in that week, negative changes representing a decrease, and 0 marking no change. The best predictive models of depression for each participant were built and selected following the steps in stage 2.

Stage 3: Encoding and Prediction Filtering Outcome

This led to stage 3, where after the mean aggregation, we encoded the regressed values as our goal was to predict whether the change in the depression score was positive, negative, or constant, rather than determining the exact value of the change. This step was followed by hamming distance calculations to further rank and filter the best set of predictions.

Stage 4: Final Prediction Ensembling of Adolescent Data

The predictions evaluated by the minimum hamming distances entered stage 4, where we calculated the final ensemble predictions. The predictions used weights determined by hamming distance calculations, which enabled us to balance between personalized and user-agnostic models. This step completed the offline training and prediction of change in depression in the adolescent data set.

Experiment Design and Results

In this section, we present the depression change prediction results of the FLMS. The experiments were designed to test the framework's claims of reducing overfitting on a small data set, reducing the impact of noise or sparsity, and identifying the best combination for sensor fusion.

We conducted 3 main experiments in support of our claims:

- Experiment 1 tested FLMS predictions against singular modeling strategies used in SOTA. This experiment evaluated our claim regarding the advantage of the overall framework that took steps to reduce noise and identify the best sensor combinations versus a singular modeling strategy.

- Experiment 2 was a SOTA comparison test conducted to evaluate how our prediction-ranking framework performed in comparison to existing ML and DL approaches used in the current literature. This comparison also substantiated the FLMS performance to overfitting versus the existing strategies in the literature from prediction in small-data scenarios.
- Experiment 3 was designed to compare the FLMS performance with that of commonly used ML algorithms that have been shown to perform well with sparse data. It is important to note that there is an overlap of ML algorithms used to tackle sparsity and those used in passive sensing studies for mental health, particularly for small data sets.

Evaluation Metrics

The task of the FLMS is to model, rank, and output the best set of predictions from multiple modeling approaches. The output of the FLMS are predictions encoded as 0s or 1s (ie, binary values). Therefore, our choice of evaluation metrics for the framework predictions was the average accuracy, average recall, and average F_1 -scores amongst users.

Experiment Metadata

The metadata pertaining to each experiment is provided at the end of the experiments. The information included as metadata is based on the best practices used [65] to help with reproducibility of results. They include (1) feature preprocessing steps, (2) modeling CV strategy, (3) ML algorithms used, (4) random state, and (5) evaluation metrics specific to the experiments. They are presented in the form of tables following the corresponding results for each experiment.

Data Set Used in the Experiments

To standardize our experiments, we maintained a consistent data set, a combination of 6-sensor feature sets that included calls, location, screen usage, conversation, Fitbit, and Wi-Fi. After the stages of preprocessing, missing data imputation using the KNN strategy, and the removal of highly correlated features, the final data set comprised 61 features and 507 data points belonging to a total of 45 (81.8%) participants.

Feature Engineering in Experiments

Since we maintained a consistent data set for all our experiments, feature engineering for all the experiments was achieved through data collected from 6 sensors. As discussed earlier, the data were collected from participants' smartphones using the AWARE app [66] and then passed through the RAPIDS application programming interface (API). The features extracted using the API are discussed in detail next.

Call Sensor Features

The calls sensor features provide a context of how frequently the user has been in contact with someone else. Studies have revealed that higher degrees of depression are linked to reduced contact with social circles [48,53]. As part of call sensor features, we extracted the total number of missed calls; the counts of missed calls from distinct contacts, calls from the most frequent contacts for a time segment, incoming calls, and outgoing calls; the mean (SD), maximum, and minimum

duration of both incoming and outgoing calls; and the entropy duration of outgoing and incoming calls, which provided an estimate of the Shannon entropy for the duration of all calls of a particular call type (ie, incoming, outgoing, or missed). All the extracted features were mean-aggregated over the period of 1 week to match the ground truth.

Location Sensor Features

Location sensor features provide a contextual idea of the amount of movement users of the sensors go through and show the correlation to mental health [3,46,49,50]. The location data are collected through the phones' GPS or the cellular towers around the phones. Location has been proven to be able to predict depressive states [3]. The features extracted from the location sensors included the location variance calculated through the sum of variance in longitude and latitude coordinates, the log of the location variance, the total distance covered, and the circadian movement [17] calculated using the Lomb-Scargle method that maps a person's location patterns following the 24-hour circadian cycle. The speed was also captured as a feature, and static labeled samples were clustered and K-means clustering was used to locate significant places visited by the participants. In addition, location entropy was also engineered to provide the proportion of time spent at each significant location visited during a day.

Screen Sensor Features

Screen sensor features are a strong indicator of how engaged users are with their phones. To capture this information, we extracted features that includes the minimum, maximum, sum, and mean (SD) of unlock episodes, along with the number of all unlock episodes and minutes until the first unlock episode. These features have been used in prior studies that proved their correlation to depressive symptom severity [46,54,55].

Conversation Sensor Features

Conversation is yet another interesting set of features that provide information pertaining to social interactions and has been used in a number of studies relating to mental health [55-58]. The computed features included the minimum, maximum, sum, and mean (SD) of the duration of all conversations. We also recorded the minutes of voice, silence, and noise. The energy associated with noise, which is the L2-norm and the sum of all energy values when noise or voice, was inferred.

Fitbit

Fitbit offers 2 features, which we extracted based on their application in previous studies relating to mental health [54,60,61], and included the maximum resting heart rate (average maximum heart rate over 1 week) and the maximum number of steps (average step count over 1 week). These features provided an idea of the physical movement and stress experienced by participants.

Wi-Fi

Wi-Fi can be a good indicator of social context. We extracted the Wi-Fi count scans that told us the number of scanned Wi-Fi access points connected to by the phone during a time segment and the number of unique connected devices during a time

segment based on the hardware address. In addition, we extracted the most scanned connected device. The use of Wi-Fi-based features in mental health prediction have been previously covered [48,59].

The data set used in our experiments had all the features discussed, which were part of the 61 features. Feature engineering helped provide a context to the data gathered from all the smartphones and Fitbit sensors and form predictions for ML models.

Results of Experiment 1

Experiment 1 showcased the overall performance of the FLMS in comparison with traditional user-agnostic and personalized models. The FLMS achieved a mean accuracy of 0.66 (SD 0.53) and a mean recall of 0.59 (SD 0.50), which are 7% and 13% higher than the best baseline performance achieved by ATU modeling. Among the singular modeling approaches, the ATU, a personalized strategy, performed best overall, with a mean accuracy of 0.59 (SD 0.50) and a mean recall of 0.46 (SD 0.66). The worst performances were shown by user-agnostic LOPO

and LTXO approaches, both of which had a mean accuracy of 0.45 (SD 0.80) and 0.47 (SD 0.83), respectively. These results are presented in Table 2 and show that singular modeling approaches used in different studies [1-4,9-17] underperform when modeling involves small, noisy, multimodal sensor data in comparison to our FLMS. The FLMS uses a balance of these strategies to improve predictions.

Experiment 1 was also designed to show how the FLMS suggests the best feature combinations for the various modeling strategies it uses through the utility of hamming distance from stage 3. The lowest hamming distance in stage 3 for the various modeling approaches used is presented in Table 3. We observed that the ATU approach led to the lowest hamming distance of 226, followed by LOTPO, with a minimum hamming distance of 267. The highest hamming distances were those of LOPO at 350 and LTXO at 378. The lower the hamming distance, the closer the predictions to ground truth. Based on this, we saw that overall, 6-sensor fusion works best for this data set. The metadata of experiment 1 are shown in Table 4.

Table 2. Experiment 1 performance of the FLMS^a in comparison to singular modeling strategies.

Modeling strategy	Type of modeling strategy	Test accuracy, mean (SD)	Test recall, mean (SD)	Test F_1 -score, mean (SD)
FLMS	User agnostic + personalized	0.66 (0.53)	0.59 (0.50)	0.56 (0.55)
ATU ^b	Personalized	0.59 (0.60)	0.46 (0.66)	0.50 (0.57)
LOTPO ^c	Personalized	0.53 (0.65)	0.45 (0.70)	0.32 (0.73)
LOPO ^d	User agnostic	0.45 (0.80)	0.43 (0.72)	0.40 (0.87)
LTXO ^e	User agnostic	0.47 (0.83)	0.35 (0.81)	0.33 (0.86)

^aFLMS: framework for longitudinal multimodal sensors.

^bATU: accumulated time unit.

^cLOTPO: leave one time unit one participant out.

^dLOPO: leave one participant out.

^eLTXO: leave time unit X out.

Table 3. Experiment 1 minimum hamming distance for choosing the best sensor combination for the experiment.

Best sensor fusion	Modeling approach in the FLMS ^a	Hamming distance
6-sensor fusion (calls + location + screen usage + conversation + Fitbit + Wi-Fi)	ATU ^b	226
6-sensor fusion (calls + location + screen usage + conversation + Fitbit + Wi-Fi)	LOTPO ^c	267
1-sensor fusion (location)	LOPO ^d	350
2-sensor fusion (calls + location)	LTXO ^e	378

^aFLMS: framework for longitudinal multimodal sensors.

^bATU: accumulated time unit.

^cLOTPO: leave one time unit one participant out.

^dLOPO: leave one participant out.

^eLTXO: leave time unit X out.

Table 4. Experiment 1 metadata.

Metadata	Experiment 1
Feature preprocessing	KNN ^a imputation, dropping highly co-related columns, sklearn StandardScaler
Modeling CV ^b strategy	FLMS ^c , ATU ^d , LOTPO ^e , LTXO ^f , LOPO ^g
ML ^h algorithms used	import XGBoost ⁱ as xgb sklearn.linear_model import LinearRegression sklearn.ensemble import RandomForestRegressor sklearn.linear_model import ElasticNet sklearn.ensemble import GradientBoostingRegressor sklearn.ensemble import ExtraTreesRegressor sklearn.ensemble import AdaBoostRegressor
Random state	42
Evaluation metrics	Accuracy, recall, F_1 -score

^aKNN: K-nearest neighbor.

^bCV: cross-validation.

^cFLMS: framework for longitudinal multimodal sensors.

^dATU: accumulated time unit.

^eLOTPO: leave one time unit one participant out.

^fLTXO: leave time unit X out.

^gLOPO: leave one participant out.

^hML: machine learning.

ⁱXGBoost: Extreme Gradient Boosting.

Results of Experiment 2

In experiment 2, we compared FLMS ranking results with ML algorithms that have been used in multiple studies on sensor-based assessment of mental health, as listed in Table 1. The ML algorithms XGBoost and KNN were chosen based on the popularity of their usage in the community, while the DL algorithm was chosen to be a basic multilayer perceptron (MLP) network and a long short-term memory (LSTM) network. These were also the best-performing algorithms compared to other ML algorithms in the literature on our data set. We initially tried using K-fold validation for the SOTA algorithms, but due to poor results, we switched to the leave-one-out strategy, which performed relatively better. This experiment first compared the overall performance of the FLMS with other SOTA algorithms based on the average test accuracy, recall, and F_1 -score. Second, the experiment substantiated the claim that the FLMS is better in tackling overfitting, as shown by the mean training accuracy versus the mean test accuracy compared to the ML algorithms in Figure 11. The models with only the single ML algorithm performed no better than the majority baseline approach, with

XGBoost showing a mean test accuracy 0.50 (SD 0.55) and the KNN showing around the same mean accuracy of 0.52 (SD 0.54), as shown in Table 5. The MLP achieved higher accuracy but a low test F_1 -score, indicating the model's performance has high false-positive and false-negative rates. The LSTM was no different and showed a similar recall and F_1 -score outcomes. The overfitting of the SOTA models is illustrated in Figure 11, where we compared the FLMS and the rest of the algorithms based on their respective performances using training and test accuracies. Figure 11 shows that the FLMS had a relatively consistent performance between a training accuracy of 68% and a test accuracy of 66%, while XGBoost, KNN, MLP, and LSTM models had high training accuracies but low test accuracies. The metadata of experiment 2 are shown in Table 6.

The experiments demonstrated support for the points highlighted in the contribution of this paper—that our ranking framework works well with small data sets in comparison to existing approaches and can reduce overfitting by using a balance-weighted ensembling of user-agnostic and personalized models.

Figure 11. Experiment 2 shows FLMS training and test accuracies in comparison to SOTA models. The FLMS is better at adapting to overfitting compared to the other algorithms. FLMS: framework for longitudinal multimodal sensors; KNN: K-nearest neighbor; LSTM: long short-term memory; ML: machine learning; MLP: multilayer perceptron; SOTA: state of the art; XGBoost: Extreme Gradient Boosting.

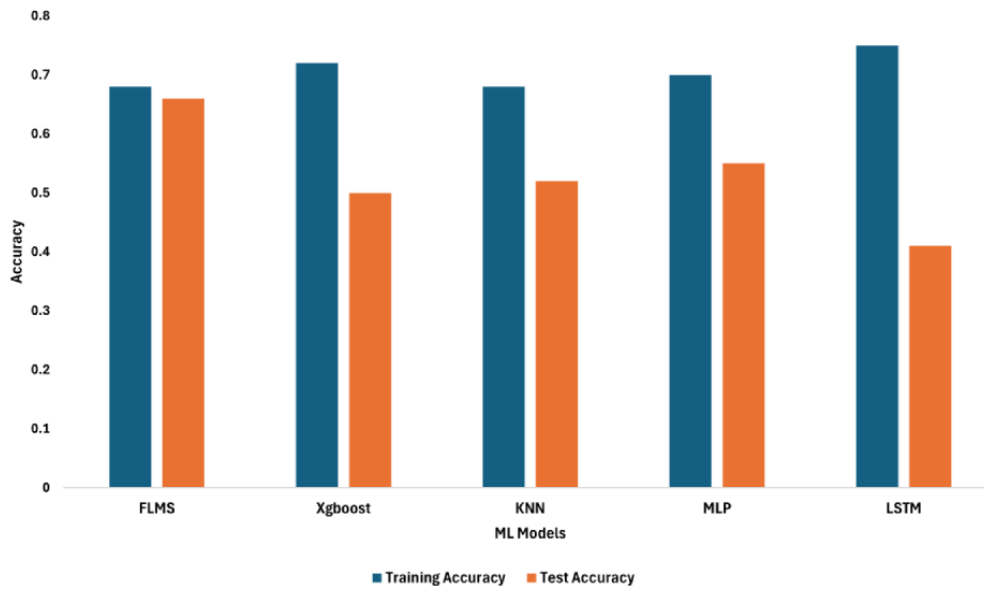


Table 5. Experiment 2 performance of the FLMS^a compared to ML^b and DL^c algorithms used in the current literature on adolescent data.

Predictive learning approach	Modeling strategy	Test accuracy, mean (SD)	Test recall, mean (SD)	Test F_1 -score, mean (SD)
FLMS	ATU ^d + LOTPO ^e + LOPO ^f + LTXO ^g	0.66 (0.53)	0.59 (0.50)	0.56 (0.55)
XGBoost ^h [14,17]	Leave 1 out	0.50 (0.55)	0.33 (0.52)	0.28 (0.57)
KNN ⁱ [10,11,13,16]	Leave 1 out	0.52 (0.54)	0.40 (0.61)	0.30 (0.73)
MLP ^j [9]	Leave 1 out	0.55 (0.70)	0.50 (0.71)	0.33 (0.70)
LSTM ^k [67]	Leave 1 out	0.41 (0.66)	0.25 (0.70)	0.35 (0.70)

^aFLMS: framework for longitudinal multimodal sensors.

^bML: machine learning.

^cDL: deep learning.

^dATU: accumulated time unit.

^eLOTPO: leave one time unit one participant out.

^fLOPO: leave one participant out.

^gLTXO: leave time unit X out.

^hXGBoost: Extreme Gradient Boosting.

ⁱKNN: K-nearest neighbor.

^jMLP: multilayer perceptron.

^kLSTM: long short-term memory.

Table 6. Experiment 2 metadata.

Metadata	Experiment 2
Feature preprocessing	KNN ^a imputation, dropping highly co-related columns, sklearn StandardScaler
Modeling CV ^b strategy	FLMS ^c , leave 1 out
ML ^d algorithms used	import XGBoost ^e as xgb sklearn.neural_network import MLPClassifier sklearn.neighbors import KNeighborsClassifier keras.layers import LSTM ^f
Random state	42
Evaluation metrics	Accuracy, recall, F_1 -Score

^aKNN: K-nearest neighbor.

^bCV: cross-validation.

^cFLMS: framework for longitudinal multimodal sensors.

^dML: machine learning.

^eXGBoost: Extreme Gradient Boosting.

^fLSTM: long short-term memory.

Results of Experiment 3

Sparsity is a challenge in dealing with small data sets. The large number of 0s or missing values can misdirect models and lead to overfitting [68]. Therefore, it is important to handle the problem of sparsity. Our experiment was designed specifically for small data sets, where sparsity proves to be a challenge. To tackle sparsity in small-data scenarios, the commonly used ML algorithms are KNN, MLP, support vector machine (SVM), decision tree (DT), random forest (RF), XGBoost, and AdaBoost [21-24,69-71].

In our experiment, we showcased a comparison of the FLMS with all the mentioned ML algorithms. We first calculated the sparsity of the adolescent data set that comprised all 6-sensor feature sets. The reason for continuing to use the 6-sensor feature sets as in the prior experiment was to test the algorithms with a data set that had a higher degree of sparsity compared to other feature combinations with lower number of sensors. The sparsity for this data set was calculated as the ratio of 0s to the total number of elements in the data set and is given as follows:



(5)

The sparsity of the data set used for this experiment was 35%. In a small data set, this is a significant amount of sparsity to negatively impact ML algorithms.

We performed the modeling and evaluated the performance based on F_1 -scores as in the case of the prediction of mental health, the F_1 -score is a good reflection of how sparsity affects the models' judgment in detecting positive and false cases. The models already shown in Table 4 remained, in addition to other models that have been mentioned in the literature to perform well on sparse data sets. Among the ML algorithms used in the literature, the best performance was shown by the RF, with an F_1 -score of 0.35, while the FLMS showed an F_1 -score 0.21 higher than that of the RF. Both MLP and AdaBoost performed close to the RF, with an F_1 -score of 0.33. The algorithm that performed the worst in handling sparsity was the SVM, with an F_1 -score of only 0.15. This experiment highlights the fact that due to the combination of modeling, the FLMS performs better when dealing with highly sparse small data sets (Table 7). The metadata of experiment 3 are shown in Table 8.

Table 7. Experiment 3 performance of the FLMS^a compared to common ML^b algorithms for tackling sparsity on the adolescent data set.

Predictive learning approach	Modeling strategy	Test F_1 -score, mean (SD)
FLMS	ATU ^c + LOTPO ^d + LOPO ^e + LTXO ^f	0.56 (0.55)
XGBoost ^g [14,17]	Leave 1 out	0.28 (0.57)
KNN ^h [10,11,13,16]	Leave 1 out	0.30 (0.73)
MLP ⁱ [9]	Leave 1 out	0.33 (0.70)
SVM ^j [12]	Leave 1 out	0.15 (0.62)
DT ^k [13]	Leave 1 out	0.24 (0.70)
RF ^l [11,13]	Leave 1 out	0.35 (0.65)
AdaBoost ^m [14]	Leave 1 out	0.33 (0.60)

^aFLMS: framework for longitudinal multimodal sensors.

^bML: machine learning.

^cATU: accumulated time unit.

^dLOTPO: leave one time unit one participant out.

^eLOPO: leave one participant out.

^fLTXO: leave time unit X out.

^gXGBoost: Extreme Gradient Boosting.

^hKNN: K-nearest neighbor.

ⁱMLP: multilayer perceptron.

^jSVM: support vector machine.

^kDT: decision tree.

^lRF: random forest.

^mAdaBoost: Adaptive Boosting.

Table 8. Experiment 3 metadata.

Metadata	Experiment 3
Feature preprocessing	KNN ^a imputation, dropping highly correlated columns, sklearn StandardScaler
Modeling CV ^b strategy	FLMS ^c , leave 1 out
ML ^d algorithms used	import XGBoost ^e as xgb from sklearn.svm import SVM ^f sklearn.neural_network import MLPClassifier sklearn.neighbors import KNeighborsClassifier sklearn.tree import DecisionTreeClassifier sklearn.ensemble import RandomForestClassifier sklearn.ensemble import AdaBoostClassifier
Random state	42
Evaluation metrics	F_1 -score

^aKNN: K-nearest neighbor.

^bCV: cross-validation.

^cFLMS: framework for longitudinal multimodal sensors.

^dML: machine learning.

^eXGBoost: Extreme Gradient Boosting.

^fSVM: support vector machine.

Discussion

Principal Findings

Solving the problem of limited and sparse data sets is not a singular modeling-based endeavor. It requires flexibility and a combination of strategies to achieve predictions that can be trusted. In this section, we discuss our ranking framework's overarching aims, performance, and limitations based on our assessments.

In experiment 1, we tested the FLMS in comparison to baseline user-agnostic and personalized models. Our framework achieved a higher accuracy, recall, and F_1 -score for the predictions when compared to singular modeling approaches, as seen in Table 2. We also demonstrated how we arrived at the sensor combination for the best set of predictions using hamming distances in stage 3 of the FLMS, as reflected in Table 3. In experiment 2, we compared the FLMS with SOTA algorithms used in the literature for predicting mental health states using sensors. The results from this experiment showed the FLMS to perform better than the existing algorithms in terms of accuracy, recall, and F_1 -scores (Table 4). Experiment 2 also highlighted the FLMS's ability to reduce overfitting in comparison to the SOTA algorithms. The FLMS showed that the training accuracy and test accuracy did not diverge by large margins, indicating it had not been overfitting the models. Lastly, we compared the FLMS ranking with that of existing ML algorithms that perform well with sparse data in experiment 3. We saw that the data set we used in our experiments exhibited 35% sparsity, which is a significant amount in an already small data set. The FLMS had a higher F_1 -score compared to the rest of the ML algorithms.

Comparison With Previous Research

The results of baseline modeling are consistent with previous studies [10,29] that showed superior performance when models were personalized. The increase in accuracy shows that our framework was able to narrow down the best set of predictions overall.

Hamming distance results showed that in LOPO and LTXO approaches, single-sensor deployment and a dual-sensor combination perform equally well as 6-sensor combinations and achieve a minimum hamming distance. This brings forth the advantage of our framework to prioritize sensor selection for yielding best predictions overall and for only the necessary number of feature sets.

The results of experiment 2 provide us with further evidence of the ranking frameworks' efficacy in balancing reliance between both user-agnostic and personalized approaches. Despite a higher accuracy, the recall of the FLMS does not overfit like that of other SOTA ML algorithms. The FLMS uses weights to balance out such effects, thus reducing the impact of overfitting in prediction performance. The test with popular existing ML algorithms showed that, despite the success of the

models in previous studies [9-11,13-17], they struggle when the data set is small and noisy, as is the case of the depression data set presented in this work. This performance result is similar when we look at the capability of ML algorithms that are better at handling sparsity. We found the FLMS to perform better than those algorithms.

Overall, seeking a single user-agnostic model that fits all is an elusive problem as most existing works suggest better performance for specialized approaches. However, specialized modeling does not perform well on heterogeneous data sets. Therefore, neither user-agnostic nor personalized modeling alone can be applicable to a specific problem area. Our framework provides a practical way to balance the 2 approaches, particularly for dealing with limited data sets.

Limitations and Future Directions

We encountered a few limitations with this study that can be addressed in future work. The FLMS was tested on the case of depression in adolescents. As such, we have not been able to establish a lower bound on the data set size that our framework is capable of handling.

Another area that we could not elaborate on is the computing speed of such a framework that might be impacted if sensor numbers rise to higher levels. Lastly, the framework was equipped with lightweight and widely used ML algorithms. Methods such as the generalized linear mixed model (GLMM) for handling longitudinal data could not be tested.

Future work can address these limitations with exposure of the framework to more multimodal, longitudinal data sets and adapting and testing other ML algorithms. Interesting future directions for the framework include its online adaptation and a similarity-based cold-start solution.

Conclusion

In this study, we presented a novel prediction-ranking framework for modeling limited noisy or sparse, multimodal, longitudinal passive sensor data. We tested our framework on an adolescent depression data set consisting of 45 participants over a period of 24 weeks. The results showed that despite the complexity and limitations of the data set, our framework is able to provide better predictions compared to singular modeling approaches. In experiment 1, our model achieved a 7% increase in accuracy and a 13% increase in recall. In experiment 2 with synthetic data, our model achieved a 5% increase in accuracy and avoided overestimating the recall value through ensembling predictions. The framework also showed its ability to explore sensor combinations through feature fusion. Our tests with existing popular SOTA algorithms showed that the models struggle when data tend to be limited and noisy. We also tested the FLMS with algorithms that perform well with sparsity and found the FLMS to exhibit a better performance. In conclusion, the FLMS can be an effective tool for passive sensing studies.

Acknowledgments

This study was supported by a grant from the National Institute of Mental Health (NIMH)(1R44MH122067); the NIMH-funded “The Center for Enhancing Triage and Utilization for Depression and Emergent Suicidality (ETUDES) in Pediatric Primary Care” (P50MH115838); the Center for Behavioral Health, Media, and Technology; and a career development award (NIMH 1K23MH11922-01A1). Research recruitment was supported by the Clinical and Translational Science Institute at the University of Pittsburgh by the National Institutes of Health Clinical and Translational Science Award (CTSA) program (grant UL1 TR001857).

Conflicts of Interest

None declared.

Multimedia Appendix 1

FLMS ranking overview. Algo: algorithm; ATU: accumulated time unit; avg: average; CV: cross-validation; FLMS: framework for longitudinal multimodal sensors; LOPO: leave one participant out; LOTPO: leave one time unit of participant out; LTXO: leave time unit X out; ML: machine learning.

[PNG File , 2313 KB - ai_v3i1e47805_app1.png]

References

1. Rabbi M, Pfammatter A, Zhang M, Spring B, Choudhury T. Automated personalized feedback for physical activity and dietary behavior change with mobile phones: a randomized controlled trial on adults. *JMIR Mhealth Uhealth* 2015 May 14;3(2):e42 [FREE Full text] [doi: [10.2196/mhealth.4160](https://doi.org/10.2196/mhealth.4160)] [Medline: [25977197](https://pubmed.ncbi.nlm.nih.gov/25977197/)]
2. Rabbi M, Aung MH, Zhang M, Choudhury T. MyBehavior: automatic personalized health feedback from user behaviors and preferences using smartphones. 2015 Presented at: UbiComp '15: 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing; September 9-11, 2015; Osaka, Japan p. 707-718 URL: <https://doi.org/10.1145/2750858.2805840> [doi: [10.1145/2750858.2805840](https://doi.org/10.1145/2750858.2805840)]
3. Canzian L, Musolesi M. Trajectories of depression: unobtrusive monitoring of depressive states by means of smartphone mobility traces analysis. 2015 Presented at: UbiComp '15: 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing; September 9-11, 2015; Osaka, Japan p. 1293-1304. [doi: [10.1145/2750858.2805845](https://doi.org/10.1145/2750858.2805845)]
4. Difrancesco S, Fraccaro P, van der Veer SN, Alshoumr B, Ainsworth J, Bellazzi R. Out-of-home activity recognition from GPS data in schizophrenic patients. 2016 Presented at: CBMS 2016: IEEE 29th International Symposium on Computer-Based Medical Systems; June 20-24, 2016; Belfast and Dublin, Ireland p. 324-328. [doi: [10.1109/cbms.2016.54](https://doi.org/10.1109/cbms.2016.54)]
5. Sano A, Phillips AJ, Amy ZY, McHill AW, Taylor S, Jaques N, et al. Recognizing academic performance, sleep quality, stress level, and mental health using personality traits, wearable sensors and mobile phones. 2015 Presented at: BSN 2015: 12th IEEE International Conference on Wearable and Implantable Body Sensor Networks; June 9-12, 2015; Cambridge, MA p. 1-6. [doi: [10.1109/bsn.2015.7299420](https://doi.org/10.1109/bsn.2015.7299420)]
6. Murahari VS, Plötz T. On attention models for human activity recognition. 2018 Presented at: ISWC '18: 2018 ACM International Symposium on Wearable Computers; October 8-12, 2018; Singapore p. 100-103 URL: <https://doi.org/10.1145/3267242.3267287> [doi: [10.1145/3267242.3267287](https://doi.org/10.1145/3267242.3267287)]
7. Allan S, Henrik B, Sourav B, Thor SP, Mikkel BK, Anind D, et al. Smart devices are different: assessing and mitigating mobile sensing heterogeneities for activity recognition. 2015 Presented at: SenSys '15: 13th ACM Conference on Embedded Networked Sensor Systems; November 1-4, 2015; Seoul, South Korea p. 127-140 URL: <https://doi.org/10.1145/2809695.2809718> [doi: [10.1145/2809695.2809718](https://doi.org/10.1145/2809695.2809718)]
8. Dietterich T. Overfitting and undercomputing in machine learning. *ACM Comput Surv* 1995 Sep;27(3):326-327. [doi: [10.1145/212094.212114](https://doi.org/10.1145/212094.212114)]
9. Kwapisz JR, Weiss GM, Moore SA. Activity recognition using cell phone accelerometers. *SIGKDD Explor Newsl* 2011 Mar 31;12(2):74-82 [FREE Full text] [doi: [10.1145/1964897.1964918](https://doi.org/10.1145/1964897.1964918)]
10. Shukla PK, Vijayvargiya A, Kumar R. Human activity recognition using accelerometer and gyroscope data from smartphones. 2020 Presented at: ICONC3: 2020 IEEE International Conference on Emerging Trends in Communication, Control and Computing; February 21-22, 2020; Lakshmangarh, Sikar, India. [doi: [10.1109/iconc345789.2020.9117456](https://doi.org/10.1109/iconc345789.2020.9117456)]
11. Chen Y, Shen C. Performance analysis of smartphone-sensor behavior for human activity recognition. *IEEE Access* 2017;5:3095-3110. [doi: [10.1109/access.2017.2676168](https://doi.org/10.1109/access.2017.2676168)]
12. Huang K, Ding X, Xu J, Guanling C, Ding W. Monitoring sleep and detecting irregular nights through unconstrained smartphone sensing. 2015 Presented at: 2015 IEEE UIC-ATC-ScalCom; August 10-14, 2015; Beijing, China p. 10-14. [doi: [10.1109/uic-atc-scalcom-cbdcom-iop.2015.30](https://doi.org/10.1109/uic-atc-scalcom-cbdcom-iop.2015.30)]
13. Montanini L, Sabino N, Spinsante S, Gambio E. Smartphone as unobtrusive sensor for real-time sleep recognition. 2018 Presented at: 2018 IEEE International Conference on Consumer Electronics (ICCE); January 12-14, 2018; Las Vegas p. 12-14 URL: <https://doi.org/10.1109/ICCE.2018.8326220> [doi: [10.1109/icce.2018.8326220](https://doi.org/10.1109/icce.2018.8326220)]

14. Teng F, Chen Y, Cheng Y, Ji X, Zhou B, Xu W. PDGes: an interpretable detection model for Parkinson's disease using smartphones. *ACM Trans Sen Netw* 2023 Apr 20;19(4):1-21 [FREE Full text] [doi: [10.1145/3585314](https://doi.org/10.1145/3585314)]
15. Azam M, Shahzadi A, Khalid A, Anwar S, Naeem U. Smartphone based human breath analysis from respiratory sounds. 2018 Presented at: 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC); July 17-21, 2018; Honolulu, HI p. 445-448. [doi: [10.1109/embc.2018.8512452](https://doi.org/10.1109/embc.2018.8512452)]
16. Grunerbl A, Osmani V, Bahle G, Carrasco JC, Oehler S, Mayora O, et al. Using smart phone mobility traces for the diagnosis of depressive and manic episodes in bipolar patients. 2014 Presented at: AH '14: 5th Augmented Human International Conference; March 7-9, 2014; Kobe, Japan p. 1-8. [doi: [10.1145/2582051.2582089](https://doi.org/10.1145/2582051.2582089)]
17. Saeb S, Lattie EG, Kording KP, Mohr DC. Mobile phone detection of semantic location and its relationship to depression and anxiety. *JMIR Mhealth Uhealth* 2017 Aug 10;5(8):e112 [FREE Full text] [doi: [10.2196/mhealth.7297](https://doi.org/10.2196/mhealth.7297)] [Medline: [28798010](https://pubmed.ncbi.nlm.nih.gov/28798010/)]
18. Plötz T. If only we had more data!: sensor-based human activity recognition in challenging scenarios. 2023 Presented at: 2023 IEEE International Conference on Pervasive Computing and Communications Workshops and other Affiliated Events (PerCom Workshops); March 13-17, 2023; Atlanta, GA p. 565-570. [doi: [10.1109/percomworkshops56833.2023.10150267](https://doi.org/10.1109/percomworkshops56833.2023.10150267)]
19. Cornet VP, Holden RJ. Systematic review of smartphone-based passive sensing for health and wellbeing. *J Biomed Inform* 2018 Jan;77:120-132 [FREE Full text] [doi: [10.1016/j.jbi.2017.12.008](https://doi.org/10.1016/j.jbi.2017.12.008)] [Medline: [29248628](https://pubmed.ncbi.nlm.nih.gov/29248628/)]
20. Xu X, Mankoff J, Dey A. Understanding practices and needs of researchers in human state modeling by passive mobile sensing. *CCF Trans Pervasive Comp Interact* 2021 Jul 06;3(4):344-366 [FREE Full text] [doi: [10.1007/s42486-021-00072-4](https://doi.org/10.1007/s42486-021-00072-4)]
21. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. 2016 Presented at: KDD '16: 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; August 13-17, 2016; San Francisco, CA p. 785-794. [doi: [10.1145/2939672.2939785](https://doi.org/10.1145/2939672.2939785)]
22. Xi Y, Xiang Z, Ramadge P, Schapire R. Speed and sparsity of regularized boosting. *PMLR* 2009;5:615-622.
23. Zou H, Hastie T. Regularization and variable selection via the elastic net. *J R Stat Soc B: Stat Methodol* 2005 Apr;67(2):301-320 [FREE Full text] [doi: [10.1111/j.1467-9868.2005.00503.x](https://doi.org/10.1111/j.1467-9868.2005.00503.x)]
24. Geurts P, Ernst D, Wehenkel L. Extremely randomized trees. *Mach Learn* 2006 Mar 2;63(1):3-42. [doi: [10.1007/s10994-006-6226-1](https://doi.org/10.1007/s10994-006-6226-1)]
25. Muhammad G, Alshehri F, Karray F, Saddik AE, Alsulaiman M, Falk TH. A comprehensive survey on multimodal medical signals fusion for smart healthcare systems. *Inf Fusion* 2021 Dec;76:355-375 [FREE Full text] [doi: [10.1016/j.inffus.2021.06.007](https://doi.org/10.1016/j.inffus.2021.06.007)]
26. Joshi S, Boyd S. Sensor selection via convex optimization. *IEEE Trans Signal Process* 2009 Feb;57(2):451-462. [doi: [10.1109/TSP.2008.2007095](https://doi.org/10.1109/TSP.2008.2007095)]
27. Altenbach F, Corroy S, Böcherer G, Mathar R. Strategies for distributed sensor selection using convex optimization. 2012 Presented at: 2012 IEEE Global Communications Conference (GLOBECOM); December 3-7, 2012; Anaheim, CA p. 2367-2372. [doi: [10.1109/glocom.2012.6503470](https://doi.org/10.1109/glocom.2012.6503470)]
28. Shorten C, Khoshgoftaar TM. A survey on image data augmentation for deep learning. *J Big Data* 2019 Jul 6;6(1):1-48 [FREE Full text] [doi: [10.1186/s40537-019-0197-0](https://doi.org/10.1186/s40537-019-0197-0)]
29. Feng SY, Gangal V, Wei J, Chandar S, Vosoughi S, Mitamura T, et al. A survey of data augmentation approaches for NLP. *arXiv*. Preprint posted online 2021. [doi: [10.48550/arXiv.2105.03075](https://doi.org/10.48550/arXiv.2105.03075)] 2021 [FREE Full text] [doi: [10.48550/arXiv.2105.03075](https://doi.org/10.48550/arXiv.2105.03075)]
30. Florez AYC, Scabora L, Amer-Yahia S, Júnior JFR. Augmentation techniques for sequential clinical data to improve deep learning prediction technique. 2020 Presented at: 2020 IEEE 33rd International Symposium on Computer-Based Medical Systems (CBMS); July 28-30, 2020; Rochester, MN p. 597-602 URL: <https://doi.org/10.1109/CBMS49503.2020.00118> [doi: [10.1109/cbms49503.2020.00118](https://doi.org/10.1109/cbms49503.2020.00118)]
31. Müller SR, Chen XL, Peters H, Chaintreau A, Matz SC. Depression predictions from GPS-based mobility do not generalize well to large demographically heterogeneous samples. *Sci Rep* 2021 Jul 07;11(1):14007 [FREE Full text] [doi: [10.1038/s41598-021-93087-x](https://doi.org/10.1038/s41598-021-93087-x)] [Medline: [34234186](https://pubmed.ncbi.nlm.nih.gov/34234186/)]
32. Xu X, Chikersal P, Dutcher JM, Sefidgar YS, Seo W, Tumminia MJ, et al. Leveraging collaborative-filtering for personalized behavior modeling: a case study of depression detection among college students. *Proc ACM Interact Mob Wearable Ubiquitous Technol* 2021 Mar 30;5(1):1-27 [FREE Full text] [doi: [10.1145/3448107](https://doi.org/10.1145/3448107)]
33. Maxhuni A, Hernandez-Leal P, Sucar LE, Osmani V, Morales EF, Mayora O. Stress modelling and prediction in presence of scarce data. *J Biomed Inform* 2016 Oct;63:344-356 [FREE Full text] [doi: [10.1016/j.jbi.2016.08.023](https://doi.org/10.1016/j.jbi.2016.08.023)] [Medline: [27592309](https://pubmed.ncbi.nlm.nih.gov/27592309/)]
34. Jacobson N, Lekkas D, Huang R, Thomas N. Deep learning paired with wearable passive sensing data predicts deterioration in anxiety disorder symptoms across 17-18 years. *J Affect Disord* 2021 Mar 01;282:104-111 [FREE Full text] [doi: [10.1016/j.jad.2020.12.086](https://doi.org/10.1016/j.jad.2020.12.086)] [Medline: [33401123](https://pubmed.ncbi.nlm.nih.gov/33401123/)]
35. Ren B, Balkind EG, Pastro B, Israel ES, Pizzagalli DA, Rahimi-Eichi H, et al. Predicting states of elevated negative affect in adolescents from smartphone sensors: a novel personalized machine learning approach. *Psychol Med* 2022 Jul 27;53(11):5146-5154. [doi: [10.1017/s0033291722002161](https://doi.org/10.1017/s0033291722002161)]

36. Adhikary A, Majumder K, Chatterjee S, Shaw RN, Ghosh A. Human activity recognition for disease detection using machine learning techniques—a comparative study. In: Shaw RN, Das S, Piuri V, Bianchini M, editors. *Advanced Computing and Intelligent Technologies. Lecture Notes in Electrical Engineering*, Vol 914. Singapore: Springer; 2022.
37. Messalas A, Kanellopoulos Y, Makris C. Model-agnostic interpretability with Shapley values. 2019 Presented at: IISA 2019: 10th IEEE International Conference on Information, Intelligence, Systems and Applications; July 15-17, 2019; Patras, Greece p. 1-7. [doi: [10.1109/iisa.2019.8900669](https://doi.org/10.1109/iisa.2019.8900669)]
38. Li L, Qiao J, Yu G, Wang L, Li HY, Liao C, et al. Interpretable tree-based ensemble model for predicting beach water quality. *Water Res* 2022 Mar 01;211:118078 [FREE Full text] [doi: [10.1016/j.watres.2022.118078](https://doi.org/10.1016/j.watres.2022.118078)] [Medline: [35066260](https://pubmed.ncbi.nlm.nih.gov/35066260/)]
39. Debie E, Fernandez Rojas R, Fidock J, Barlow M, Kasmarik K, Anavatti S, et al. Multimodal fusion for objective assessment of cognitive workload: a review. *IEEE Trans Cybern* 2021 Mar;51(3):1542-1555. [doi: [10.1109/tcyb.2019.2939399](https://doi.org/10.1109/tcyb.2019.2939399)]
40. Pedrelli P, Fedor S, Ghandeharioun A, Howe E, Ionescu DF, Bhatena D, et al. Monitoring changes in depression severity using wearable and mobile sensors. *Front Psychiatry* 2020 Dec 18;11:584711 [FREE Full text] [doi: [10.3389/fpsy.2020.584711](https://doi.org/10.3389/fpsy.2020.584711)] [Medline: [33391050](https://pubmed.ncbi.nlm.nih.gov/33391050/)]
41. Wang R. On predicting relapse in schizophrenia using mobile sensing in a randomized control trial. 2020 Presented at: 2020 IEEE International Conference on Pervasive Computing and Communications (PerCom); March 23-27, 2020; Austin, TX p. 1-8. [doi: [10.1109/percom45495.2020.9127365](https://doi.org/10.1109/percom45495.2020.9127365)]
42. Sun S, Folarin AA, Zhang Y, Cummins N, Garcia-Dias R, Stewart C, RADAR-CNS Consortium. Challenges in using mHealth data from smartphones and wearable devices to predict depression symptom severity: retrospective analysis. *J Med Internet Res* 2023 Aug 14;25:e45233 [FREE Full text] [doi: [10.2196/45233](https://doi.org/10.2196/45233)] [Medline: [37578823](https://pubmed.ncbi.nlm.nih.gov/37578823/)]
43. Tlachac M, Toto E, Lovering J, Kayastha R, Taurich N, Rundensteiner E. EMU: early mental health uncovering framework and dataset. 2021 Presented at: ICMLA 2021: 20th IEEE International Conference on Machine Learning and Applications; December 13-16, 2021; Pasadena, CA p. 1311-1318. [doi: [10.1109/icmla52953.2021.00213](https://doi.org/10.1109/icmla52953.2021.00213)]
44. Negeri ZF, Levis B, Sun Y, He C, Krishnan A, Wu Y, Depression Screening Data (DEPRESSD) PHQ Group. Accuracy of the Patient Health Questionnaire-9 for screening to detect major depression: updated systematic review and individual participant data meta-analysis. *BMJ* 2021 Oct 05;375:n2183 [FREE Full text] [doi: [10.1136/bmj.n2183](https://doi.org/10.1136/bmj.n2183)] [Medline: [34610915](https://pubmed.ncbi.nlm.nih.gov/34610915/)]
45. Vega J, Li M, Aguilera K, Goel N, Joshi E, Khandekar K, et al. Reproducible analysis pipeline for data streams: open-source software to process data collected with mobile devices. *Front Digit Health* 2021 Nov 18;3:769823 [FREE Full text] [doi: [10.3389/fdgh.2021.769823](https://doi.org/10.3389/fdgh.2021.769823)] [Medline: [34870271](https://pubmed.ncbi.nlm.nih.gov/34870271/)]
46. Saeb S, Zhang M, Karr CJ, Schueller SM, Corden ME, Kording KP, et al. Mobile phone sensor correlates of depressive symptom severity in daily-life behavior: an exploratory study. *J Med Internet Res* 2015 Jul 15;17(7):e175 [FREE Full text] [doi: [10.2196/jmir.4273](https://doi.org/10.2196/jmir.4273)] [Medline: [26180009](https://pubmed.ncbi.nlm.nih.gov/26180009/)]
47. Wang R, Wang W, daSilva A, Huckins JF, Kelley WM, Heatherton TF, et al. Tracking depression dynamics in college students using mobile phone and wearable sensing. *Proc ACM Interact Mob Wearable Ubiquitous Technol* 2018 Mar 26;2(1):1-26. [doi: [10.1145/3191775](https://doi.org/10.1145/3191775)]
48. Wahle F, Kowatsch T, Fleisch E, Rufer M, Weidt S. Mobile sensing and support for people with depression: a pilot trial in the wild. *JMIR Mhealth Uhealth* 2016 Sep 21;4(3):e111,e5960 [FREE Full text] [doi: [10.2196/mhealth.5960](https://doi.org/10.2196/mhealth.5960)] [Medline: [27655245](https://pubmed.ncbi.nlm.nih.gov/27655245/)]
49. Mehrotra A, Musolesi M. Using autoencoders to automatically extract mobility features for predicting depressive states. *Proc ACM Interact Mob Wearable Ubiquitous Technol* 2018 Sep 18;2(3):1-20. [doi: [10.1145/3264937](https://doi.org/10.1145/3264937)]
50. Farhan AA, Yue C, Morillo R, Ware S, Lu J, Bi J, et al. Behavior vs. introspection: refining prediction of clinical depression via smartphone sensing data. 2016 Presented at: 2016 IEEE Wireless Health; October 25-27, 2016; Bethesda, MD p. 30-37. [doi: [10.1109/wh.2016.7764553](https://doi.org/10.1109/wh.2016.7764553)]
51. Chikersal P, Doryab A, Tumminia M, Villalba DK, Dutcher JM, Liu X, et al. Detecting depression and predicting its onset using longitudinal symptoms captured by passive sensing: a machine learning approach with robust feature selection. *ACM Trans Comput-Hum Interact* 2021 Jan 20;28(1):1-41. [doi: [10.1145/3422821](https://doi.org/10.1145/3422821)]
52. Lane ND, Lin M, Mohammad M, Yang X, Lu H, Cardone G, et al. BeWell: sensing sleep, physical activities and social interactions to promote wellbeing. *Mobile Netw Appl* 2014 Jan 9;19(3):345-359 [FREE Full text] [doi: [10.1007/s11036-013-0484-5](https://doi.org/10.1007/s11036-013-0484-5)]
53. LiKamWa R, Liu Y, Lane N, Zhong L. MoodScope: building a mood sensor from smartphone usage patterns. 2013 Presented at: MobiSys'13: 11th Annual International Conference on Mobile Systems, Applications, and Services; June 25-28, 2013; Taipei, Taiwan p. 25-28. [doi: [10.1145/2462456.2464449](https://doi.org/10.1145/2462456.2464449)]
54. Doryab A, Villalba DK, Chikersal P, Dutcher JM, Tumminia M, Liu X, et al. Identifying behavioral phenotypes of loneliness and social isolation with passive sensing: statistical analysis, data mining and machine learning of smartphone and Fitbit data. *JMIR Mhealth Uhealth* 2019 Jul 24;7(7):e13209 [FREE Full text] [doi: [10.2196/13209](https://doi.org/10.2196/13209)] [Medline: [31342903](https://pubmed.ncbi.nlm.nih.gov/31342903/)]
55. Wang R, Aung MSH, Abdullah S. CrossCheck: toward passive sensing and detection of mental health changes in people with schizophrenia. 2016 Presented at: UbiComp '16: 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing; September 12-16, 2016; Heidelberg, Germany. [doi: [10.1145/2971648.2971740](https://doi.org/10.1145/2971648.2971740)]

56. Lane N, Rabbi M, Lin M, Yang X. Bewell: a smartphone application to monitor, model and promote wellbeing. 2012 Presented at: 5th International ICST Conference on Pervasive Computing Technologies for Healthcare; May 23-26, 2011; Dublin, Ireland. [doi: [10.4108/icst.pervasivehealth.2011.246161](https://doi.org/10.4108/icst.pervasivehealth.2011.246161)]
57. Mashfiqui R, Ali S, Choudhury T, Berke E. Passive and in-situ assessment of mental and physical well-being using mobile sensors. 2011 Presented at: UbiComp '11: 13th International Conference on Ubiquitous Computing; September 17-21, 2011; Beijing, China p. 385-394. [doi: [10.1145/2030112.2030164](https://doi.org/10.1145/2030112.2030164)]
58. Wang R, Chen F, Chen Z. StudentLife: assessing mental health, academic performance and behavioral trends of college students using smartphones. 2014 Presented at: UbiComp '14: 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing; September 13-17, 2014; Seattle, WA p. 3-14. [doi: [10.1145/2632048.2632054](https://doi.org/10.1145/2632048.2632054)]
59. Ware S, Yue C, Morillo R, Lu J, Shang C, Kamath J, et al. Large-scale automatic depression screening using meta-data from WiFi infrastructure. Proc ACM Interact Mob Wearable Ubiquitous Technol 2018 Dec 27;2(4):1-27. [doi: [10.1145/3287073](https://doi.org/10.1145/3287073)]
60. Dai R, Kannampallil T, Kim S. Detecting mental disorders with wearables: a large cohort study. 2023 Presented at: IoTDI '23: 8th ACM/IEEE Conference on Internet of Things Design and Implementation; May 9-12, 2023; San Antonio, TX p. 39-51. [doi: [10.1145/3576842.3582389](https://doi.org/10.1145/3576842.3582389)]
61. Doryab A, Chikarsel P, Liu X, Dey AK. Extraction of behavioral features from smartphone and wearable data. arXiv. Preprint posted online 2018. [doi: [10.48550/arXiv.1812.10394](https://doi.org/10.48550/arXiv.1812.10394)] 2021. [doi: [10.48550/arXiv.1812.10394](https://doi.org/10.48550/arXiv.1812.10394)]
62. Kroenke K, Strine TW, Spitzer RL, Williams JB, Berry JT, Mokdad AH. The PHQ-8 as a measure of current depression in the general population. J Affect Disord 2009 Apr;114(1-3):163-173. [doi: [10.1016/j.jad.2008.06.026](https://doi.org/10.1016/j.jad.2008.06.026)] [Medline: [18752852](https://pubmed.ncbi.nlm.nih.gov/18752852/)]
63. Hamilton M. The Hamilton Rating Scale for depression. In: Assessment of Depression. Berlin, Heidelberg: Springer; 1986:143-152.
64. Thompson E. Hamilton Rating Scale for anxiety (HAM-A). Occup Med (Lond) 2015 Oct 13;65(7):601. [doi: [10.1093/occmed/kqv054](https://doi.org/10.1093/occmed/kqv054)] [Medline: [26370845](https://pubmed.ncbi.nlm.nih.gov/26370845/)]
65. Schelter S, Böse JH, Kirschnick J, Klein T, Seufert S. Automatically tracking metadata and provenance of machine learning experiments. Amazon Science. 2017. URL: <https://assets.amazon.science/2f/39/4b32cf354e4c993b439d88258597/automaticaly-tracking-metadata-and-provenance-of-machine-learning-experiments.pdf> [accessed 2024-05-01]
66. Ferreira D, Kostakos V, Dey AK. AWARE: mobile context instrumentation framework. Front ICT 2015 Apr 20;2:6. [doi: [10.3389/fict.2015.00006](https://doi.org/10.3389/fict.2015.00006)]
67. Acikmese Y, Alptekin SE. Prediction of stress levels with LSTM and passive mobile sensors. Procedia Comput Sci 2019;159:658-667. [doi: [10.1016/j.procs.2019.09.221](https://doi.org/10.1016/j.procs.2019.09.221)]
68. Kucukozer-Cavdar S, Taskaya-Temizel T, Mehrotra A, Musolesi M, Tino P. Designing robust models for behaviour prediction using sparse data from mobile sensing: a case study of office workers' availability for well-being interventions. ACM Trans Comput Healthc 2021 Jul 18;2(4):1-33 [FREE Full text] [doi: [10.1145/3458753](https://doi.org/10.1145/3458753)]
69. Yin D, Li J, Wu G. Solving the data sparsity problem in predicting the success of the startups with machine learning methods. arXiv. Preprint posted online 2021. [doi: [10.48550/arXiv.2112.07985](https://doi.org/10.48550/arXiv.2112.07985)] 2021;07985(2021). [doi: [10.48550/arXiv.2112.07985](https://doi.org/10.48550/arXiv.2112.07985)]
70. Zhang M, Sun Y, Liang F. Sparse deep learning for time series data: theory and applications. arXiv. Preprint posted online 2023. [doi: [10.48550/arXiv.2310.03243](https://doi.org/10.48550/arXiv.2310.03243)] 2021. [doi: [10.48550/arXiv.2310.03243](https://doi.org/10.48550/arXiv.2310.03243)]
71. Rosidi N. Best machine learning model for sparse data. KD nuggets. 2023 Apr 7. URL: <https://www.kdnuggets.com/2023/04/best-machine-learning-model-sparse-data.html> [accessed 2024-05-01]

Abbreviations

- AdaBoost:** Adaptive Boosting
- API:** application programming interface
- ATU:** accumulated time unit
- CV:** cross-validation
- DL:** deep learning
- DT:** decision tree
- FLMS:** framework for longitudinal multimodal sensors
- HAR:** human activity recognition
- KNN:** K-nearest neighbor
- LDA:** linear discriminant analysis
- LOPO:** leave one participant out
- LOTPO:** leave one time unit one participant out
- LR:** linear regression
- LSTM:** long short-term memory
- LTXO:** leave time unit X out
- ML:** machine learning
- MLP:** multilayer perceptron

PCA: principal component analysis
PHQ-9: 9-item Patient Health Questionnaire
RF: random forest
SOTA: state of the art
SVM: support vector machine
XGBoost: Extreme Gradient Boosting

Edited by Y Huo; submitted 02.04.23; peer-reviewed by L Zheng, A Tomar; comments to author 02.07.23; revised version received 16.09.23; accepted 09.04.24; published 20.05.24.

Please cite as:

Mullick T, Shaaban S, Radovic A, Doryab A

Framework for Ranking Machine Learning Predictions of Limited, Multimodal, and Longitudinal Behavioral Passive Sensing Data: Combining User-Agnostic and Personalized Modeling

JMIR AI 2024;3:e47805

URL: <https://ai.jmir.org/2024/1/e47805>

doi: [10.2196/47805](https://doi.org/10.2196/47805)

PMID: [38875667](https://pubmed.ncbi.nlm.nih.gov/38875667/)

©Tahsin Mullick, Sam Shaaban, Ana Radovic, Afsaneh Doryab. Originally published in JMIR AI (<https://ai.jmir.org>), 20.05.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR AI, is properly cited. The complete bibliographic information, a link to the original publication on <https://www.ai.jmir.org/>, as well as this copyright and license information must be included.

Original Paper

Predicting Workers' Stress: Application of a High-Performance Algorithm Using Working-Style Characteristics

Hiroki Iwamoto¹, MSc; Saki Nakano¹, MEng; Ryotaro Tajima¹, MSI; Ryo Kiguchi¹, MSc; Yuki Yoshida¹, MSc; Yoshitake Kitanishi¹, PhD; Yasunori Aoki^{1,2}, MD, PhD

¹Shionogi & Co., Ltd., Osaka, Japan

²Department of Psychiatry, Nippon Life Hospital, Osaka, Japan

Corresponding Author:

Hiroki Iwamoto, MSc

Shionogi & Co., Ltd.

Awajimachi Office 4F, Midousuji MTR Building

6-3 Awajimachi 3-chome, Chuo-ku

Osaka, 541-0047

Japan

Phone: 81 90 9540 3570

Email: hiroki.iwamoto@shionogi.co.jp

Abstract

Background: Work characteristics, such as teleworking rate, have been studied in relation to stress. However, the use of work-related data to improve a high-performance stress prediction model that suits an individual's lifestyle has not been evaluated.

Objective: This study aims to develop a novel, high-performance algorithm to predict an employee's stress among a group of employees with similar working characteristics.

Methods: This prospective observational study evaluated participants' responses to web-based questionnaires, including attendance records and data collected using a wearable device. Data spanning 12 weeks (between January 17, 2022, and April 10, 2022) were collected from 194 Shionogi Group employees. Participants wore the Fitbit Charge 4 wearable device, which collected data on daily sleep, activity, and heart rate. Daily work shift data included details of working hours. Weekly questionnaire responses included the K6 questionnaire for depression/anxiety, a behavioral questionnaire, and the number of days lunch was missed. The proposed prediction model used a neighborhood cluster (N=20) with working-style characteristics similar to those of the prediction target person. Data from the previous week predicted stress levels the following week. Three models were compared by selecting appropriate training data: (1) single model, (2) proposed method 1, and (3) proposed method 2. Shapley Additive Explanations (SHAP) were calculated for the top 10 extracted features from the Extreme Gradient Boosting (XGBoost) model to evaluate the amount and contribution direction categorized by teleworking rates (mean): low: <0.2 (more than 4 days/week in office), middle: 0.2 to <0.6 (2 to 4 days/week in office), and high: ≥0.6 (less than 2 days/week in office).

Results: Data from 190 participants were used, with a teleworking rate ranging from 0% to 79%. The area under the curve (AUC) of the proposed method 2 was 0.84 (true positive vs false positive: 0.77 vs 0.26). Among participants with low teleworking rates, most features extracted were related to sleep, followed by activity and work. Among participants with high teleworking rates, most features were related to activity, followed by sleep and work. SHAP analysis showed that for participants with high teleworking rates, skipping lunch, working more/less than scheduled, higher fluctuations in heart rate, and lower mean sleep duration contributed to stress. In participants with low teleworking rates, coming too early or late to work (before/after 9 AM), a higher/lower than mean heart rate, lower fluctuations in heart rate, and burning more/fewer calories than normal contributed to stress.

Conclusions: Forming a neighborhood cluster with similar working styles based on teleworking rates and using it as training data improved the prediction performance. The validity of the neighborhood cluster approach is indicated by differences in the contributing features and their contribution directions among teleworking levels.

Trial Registration: UMIN UMIN000046394; <https://www.umin.ac.jp/ctr/index.htm>

(JMIR AI 2024;3:e55840) doi:[10.2196/55840](https://doi.org/10.2196/55840)

KEYWORDS

high-performance algorithm; Japan; questionnaire; stress prediction model; teleworking; wearable device

Introduction

Stress is an external or internal stimulus that produces a compensatory biological response that can trigger or aggravate many diseases or pathological conditions [1]. Notably, the stress-depression association requires recognizing the effects of context and personal characteristics on the existence of stressors and understanding the progressive and dynamic relationship between stress and depression over time [2]. This is important because depression remains a major social issue [3] with a high relapse rate, prolonged duration of illness [4], and high socioeconomic impact [5]. The duration of untreated depression is associated with worse outcomes [6]. The annual national cost of major depressive disorder among adults aged ≥ 20 years in Japan in 2008 was approximately US \$11 billion, including US \$6.9 billion in workplace-associated expenses [5].

Detecting and targeting depression before a formal diagnosis can serve as an early countermeasure to depression. Therefore, detecting stress in advance is vital because stress is a factor that triggers depression and increases the risk of relapse [2]. Companies are placing an ever-increasing emphasis on their employees' mental health, including their experience of stress, as an important topic to address. According to the Japanese Ministry of Health, Labour and Welfare (2021), the proportion of companies with workers taking temporary leave or retiring due to mental health conditions has increased from 9.2% in 2020 to 10.1% in 2021 [7]. Furthermore, about 40% of companies in Japan reported worsening employee mental health due to the COVID-19 pandemic [8]. Therefore, in response to this growing need, the proportion of companies conducting stress checks on their employees has increased from 62.7% in 2020 to 65.2% in 2021 in Japan [7].

One approach is to develop stress prediction models using data related to stress collected by wearable devices that measure parameters such as heart rate variability [9], physical activity [10], and sleep [11], as well as through questionnaire responses that provide insights into physical activity [12] (eg, outings), absenteeism (failure to report for scheduled work), and the number of times lunch is missed [13]. However, these data are affected by working style such as teleworking habits (eg, remote working).

To the best of our knowledge, there is no study taking teleworking habits into account for stress prediction even though the relationship between teleworking and stress has been studied. Teleworking/telecommuting can have an impact on mental health [14,15]. However, stress is dependent not only on the environment but also on an individual's attributes [16,17]. Moreover, stress parameters [9,18,19] can be influenced by various other factors. Consequently, a few studies on stress detection have used a personalized model-based approach [20-22].

The objective of this study was to develop a novel, high-performance stress prediction algorithm using working data focusing on employees' teleworking habits.

Methods

Study Design

This prospective observational study (UMIN000046394) evaluated participants' responses to web-based questionnaires, including attendance records and data collected via a wearable device. The data were used to develop a high-performance stress prediction algorithm based on working-style characteristics similar to those of the prediction target person among the participants. Data spanning 12 weeks were collected for each employee from January 17, 2022, to April 10, 2022.

Ethical Considerations

Informed consent was obtained from employees using a web-based consent form. This study was approved by the Research and Ethics Committee of Shionogi & Co., Ltd (EP21-13) and the MINS Institutional Review Board (210238), a specified nonprofit organization. The study was conducted in compliance with the ethical guidelines for medical and health research involving human participants and in accordance with the ethical principles of the Declaration of Helsinki. To deidentify the participants, age and sex data were not collected.

Recruitment

This study enrolled 194 employees of the Shionogi Group working in Osaka, Japan. Participants who rarely teleworked included sales or research employees, and those who frequently teleworked included clerical employees. Notably, neither 100% teleworking nor teleworking other than working from home was permitted for Shionogi Group employees. The teleworking rate was calculated as the number of days an employee worked from home during the 12 weeks divided by the number of days an employee worked during the 12 weeks.

The participants, who were from different departments, worked during standard working hours (9 AM to 5 AM Monday to Friday); however, given the anticipated flexible time system for data collection, participants could decide their working hours each day and enter work start and end times into the attendance management system in advance. Night shift workers were not included in this study, and while there was a certain degree of flexibility in work hours, daytime workers were encouraged not to shift their work hours too far from the standard workday except when necessary. There were no exclusion criteria other than working time and region (daytime employee, working in Osaka), thereby reducing enrollment bias.

Data Collection

Daily data collected from the Fitbit Charge4 wearable device worn for 12 weeks (Fitbit LLC) included sleep data recorded daily (sleep duration, sleep efficiency, sleep initiation, and end time), activity data recorded every 15 min (number of steps taken, distance moved, number of floors climbed or descended, and calories burned), and heart rate per minute. Daily work shift data collected included working hours, scheduled work start

and end times, scheduled hours of work, work from home (yes/no), and absence from work/leave taken (yes/no).

Weekly web-based questionnaire responses included the K6 questionnaire [23,24], which measures 6 common symptoms of depression and anxiety, each rated on a scale between 0 and 4 (0=never, 1=a little, 2=sometimes, 3=most often, and 4=at all times). The total score was the sum of the responses to each question (ranging from 0 to 24), the behavioral questionnaire (number of outings, such as commuting and social outings), and the number of days lunch was missed. We selected the latter 2 parameters based on the premise that the number of outings is an alternative index for exercise habits [12]. Outings could also be used as an alternative index for UV exposure, which is reported to be related to mental health [25,26], and skipping lunch is reported to be related to stress [13].

Proposed Prediction Model

Step 1: Extract the Neighborhood Cluster

The participants were arranged in ascending order based on their teleworking rate, with each participant serving as a prediction target person. To homogenize the training data background, a group of participants whose working style/work characteristics were similar to those of the prediction target person were extracted and labeled as the neighborhood cluster.

This neighborhood cluster included participants with the top 20 nearest teleworking rates (for the training data) from the prediction target person. In some instances, when the size of the neighborhood cluster was greater than 20 because of the same ranking on the boundary, participants on the boundary were randomly sampled to include only 20 participants.

Step 2: Create an Individual Model to Predict Stress

The selected neighborhood cluster was subsequently used to train a prediction model for each prediction target person, meaning that an “n” number of different prediction models was created for the “n” number of targets to be included in this analysis. Using the neighborhood cluster data extracted in Step 1, a model was created that was individually optimized for the prediction target person. Data from the previous week were used to predict the stress level in the following week using this individual model. Although data for 12 weeks were collected, only the data for 11 weeks were used in the model because the data before week 1 (-1 week) were not collected to use the first-week data in the model (Figure 1).

The 12-week data were split into training and test data for the 3 models. The training data comprised all 12-week data of the neighborhood cluster plus data from the first 7 weeks for the prediction target person. The test data comprised the last 5 weeks of data from the prediction target person (Figure 2).

Figure 1. Prediction model. Data collected within a term shown by a blue dashed-line box are input to the prediction model, and the stress state (negative/positive) at the timepoint shown by a red star is predicted.

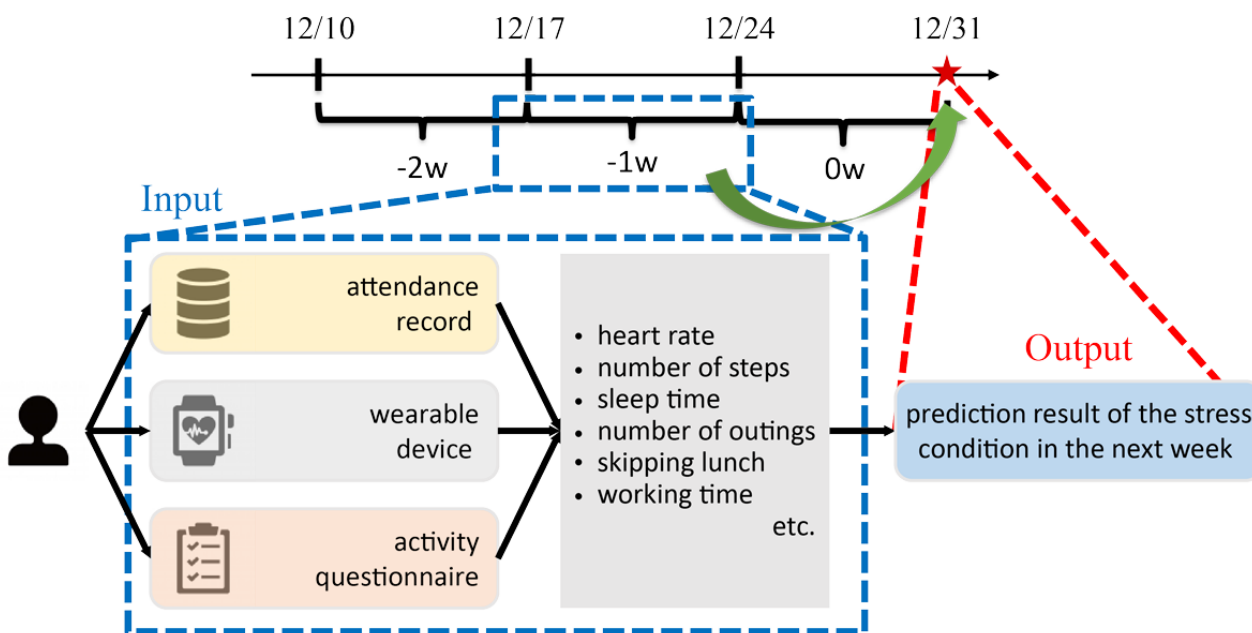
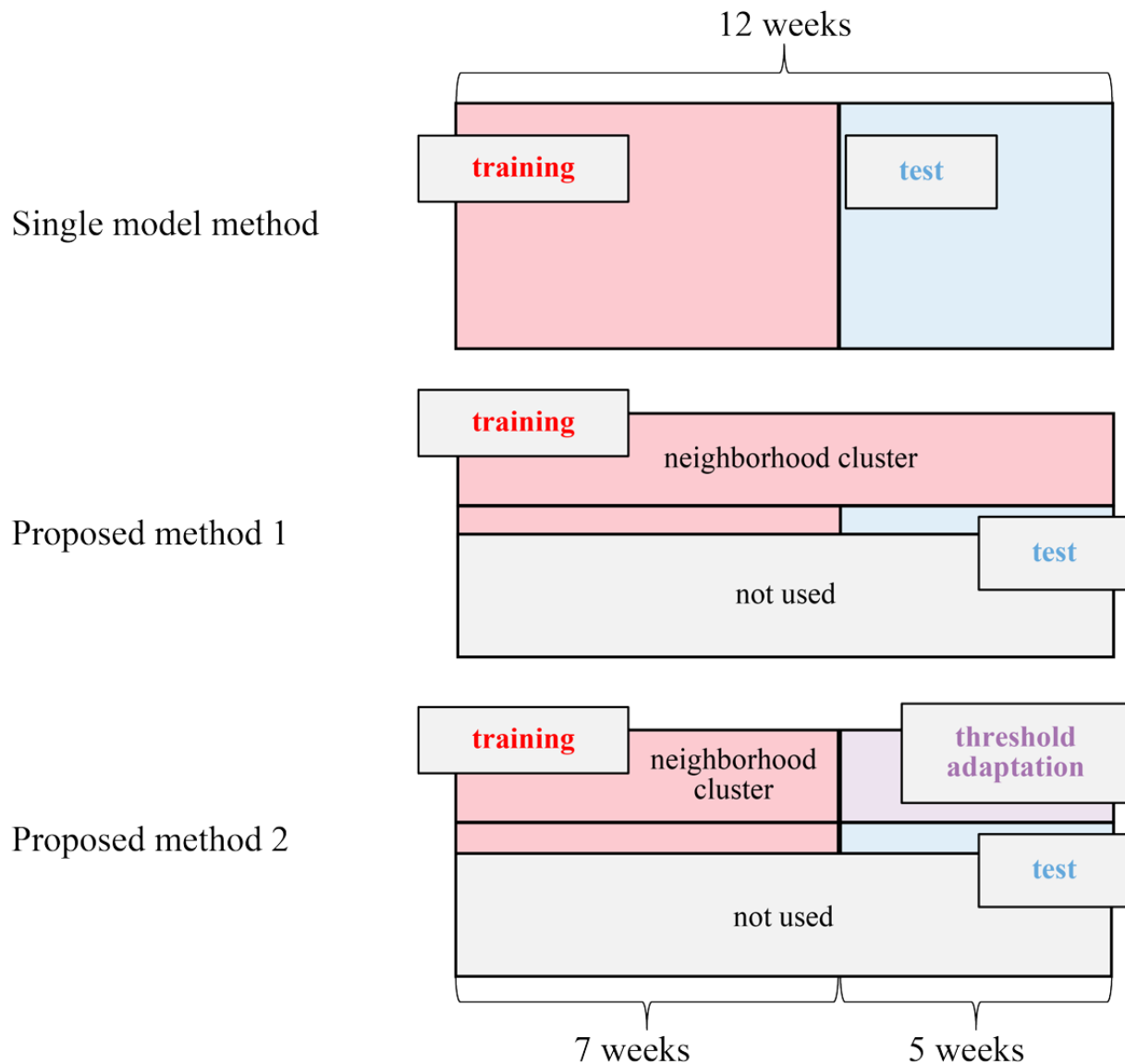


Figure 2. Twelve-week data split for comparison of the three methods.



Analysis Method

Sample Size

Considering the feasibility of an exploratory evaluation, the number of study participants was set to 150. However, the proportion of people with mental illness at the Shionogi Group was estimated to be between 7% and 10%, and the expected participation of approximately 10 patients with mental illness was based on this value. In general, too few mental illness cases lead to failure of analysis, whereas too many mental illness cases (>10%) do not appropriately reflect the population. As a screening method, we collected a stress check questionnaire when obtaining informed consent. However, as the number of mental illness cases was within the expected range of 7% to 10%, a formal screening was not performed. A total of 2037 weeks of data were evaluated. Data were evaluated weekly, and the mean (SD) was calculated from each participant's weekly data. The mean was omitted only when data were missing for the entire 7 days of the week, and the SD was omitted only when data were missing for ≥ 6 days of the week (unbiased SD

required 2 or more data points). The K6 questionnaire scores representing the stress index [23] were converted into binary objective variables (negative=K6: 0-4 [class 1]; positive=K6: 5-8 [class 2], K6: 9-12 [class 3], and K6: ≥ 13 [class 4]).

Model Training Details

The analysis was performed using Python (version 3.8.0; Python Software Foundation) and PyCaret (version 2.3.10). The Extreme Gradient Boosting (XGBoost) hyperparameters were set as follows (common in all cases): `max_depth=6`, `learning_rate=0.3`, and `n_estimators=100`. These hyperparameter values are the default configuration of PyCaret, and a hyperparameter search was not performed. The 3 models were compared, which included threshold adaptation. The single model used the first 7 weeks as training data and the latter 5 weeks as test data for all participants. Proposed method 1 used 12-week data of the neighborhood cluster plus the first 7-week data of the prediction target person as training data and the latter 5-week data of the prediction target person as test data. Both methods used a fixed threshold of 0.5 (the default threshold of XGBoost); an output of the stress prediction model above this

threshold indicated high levels of stress. Proposed method 2 used 7-week data of the neighborhood cluster and the prediction target person as training data, the latter 5-week data of the prediction target person as test data, and the latter 5-week of the neighborhood cluster for threshold adaptation. The explanatory variables are the 50 features shown in [Multimedia Appendix 1](#), and the object variable is the binarized stress score.

The threshold was adjusted such that the true positive (TP) rate was >0.8 using the threshold adaptation data. A value of 0.8 was the practically required TP rate. Of note, there was no guarantee that the TP rate would be >0.8 in the test data because the threshold was not adjusted for test data. The prediction threshold was adjusted such that the TP rate increased to >0.8 , with the lowest false positive (FP) rate. Notably, determining the TP rate is more important than determining the FP rate to ensure early depression countermeasures. Thus, by setting the value to 0.8, we could predict as many positives as possible. The area under the receiver operating characteristic curve (AUROC) was used to measure the performance of the models.

Data Exclusion

A total of 190 individual models were created, as 2 participants discontinued the study, and data from 2 other participants were missing in the latter 5 weeks and were not included in the test data. However, the data of the latter 2 participants were available for the first 7 weeks and were thus included in the training data ([Figure 2](#)).

Procedure for Checking Feature Contribution

We selected figures to report the absolute amount of feature contribution and feature contribution variability between teleworking rates. Feature importance for the prediction was evaluated for each individual model using XGBoost [27,28], and the top 10 features were identified. High feature importance was defined as the factor (50 variables shown in [Multimedia Appendix 1](#)) with a high contribution (influence) to the prediction. Feature importance was defined as a score calculated based on the reduction in the objective function related to heterogeneity (sum of squared residuals for continuous variables and the Gini index for categorical variables) achieved by

splitting the feature value when creating decision trees ([Multimedia Appendix 2](#)) [28].

Thereafter, the individual model was divided into 3 levels stratified by the teleworking rate, and the top 10 feature values for each level were extracted. Finally, Shapley Additive Explanations (SHAP) [29] were calculated for the top 10 extracted features to evaluate their impact and contribution direction, stratified by 3 levels of teleworking rates, as follows: (1) low: <0.2 (mean of >4 days per week in office), (2) middle: 0.2 to <0.6 (mean of 2-4 days per week in office), and (3) high: ≥ 0.6 (mean of <2 days per week in office). The absolute value of SHAP represents the contribution amount, while its positive or negative direction on the y-axis represents the contribution direction.

The contribution direction and impact of features were based on “covariance of features and SHAP” divided by “SD of features.” Any positive deviation from 0 on the y-axis was considered to positively impact stress, and any negative deviation was considered to negatively impact stress.

Results

Overall Findings

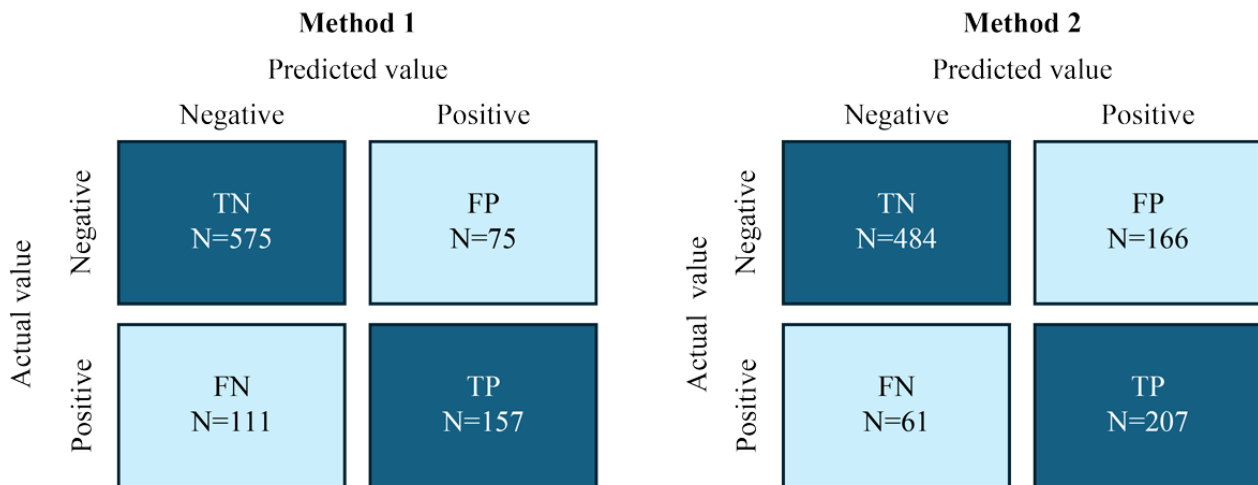
Data from 190/194 (97.9%) participants were included to develop high-performance stress prediction algorithms; 2 participants discontinued the study, and data from 2 other participants were included only in the training set. The teleworking rate of the employees ranged between 0% and 79%. The prediction results of the individual models were integrated for all participants using proposed methods 1 and 2 and compared with the results of the single model. Although the proposed methods improved the prediction performance, the AUC was similar for proposed methods 1 and 2. The AUC was the highest for proposed method 1, at 0.85 (TP vs FP: 0.59 vs 0.12), followed by proposed method 2, at 0.84 (TP vs FP: 0.77 vs 0.26) and the single model method, at 0.76 (TP vs FP: 0.42 vs 0.12) ([Table 1](#)). The confusion matrix for methods 1 and 2 is presented in [Figure 3](#).

Table 1. Comparison of prediction results of the single model method and proposed methods 1 and 2.

Performance metric	Single model	Proposed method 1	Proposed method 2
True positive rate	0.42	0.59	0.77
False positive rate	0.12	0.12	0.26
AUROC ^a	0.76	0.85	0.84

^aAUROC: area under the receiver operating characteristic curve.

Figure 3. Confusion matrix for methods 1 and 2. “N” represents the total number of target classes. FN: false negative; FP: false positive; TN: true negative; TP: true positive.



Feature Importance Analysis

The top 10 features with the highest mean feature importance ranking for each of the 3 teleworking levels are presented in Multimedia Appendix 2. These 10 features were divided into 3 categories: activity (red), work (green), and sleep (blue). They were then tabulated by teleworking levels, with 43.2% (n=82) at the low level, 36.3% (n=69) at the middle level, and 20.5% (n=39) at the high level. Among the participants with a low teleworking rate, most features were related to sleep, followed by activity and work. Among the participants with high teleworking rates, most features were related to activity, followed by sleep and work.

Analysis of Feature Contribution Direction Based on SHAP

The contribution direction of each individual model for the top 10 extracted features was examined at each level. Although

many features were evaluated, only those with interesting suggestions have been reported. Middle and low teleworking rates and longer working hours contributed to higher stress levels (Figure 4A). Irrespective of the teleworking rate, lower activity contributed to higher stress levels (Figure 4B).

Participants with a high teleworking rate who skipped lunch more often had higher stress levels than those with low or middle teleworking rates. Interestingly, skipping lunch did not contribute to stress prediction in participants with middle and low teleworking rates (Figure 5A). Working more or less than scheduled hours (high variation in the working hour gap) contributed to stress, especially for those with high teleworking rates (Figure 5B). Low fluctuations in heart rate (SD of the heart rate) contributed to stress, particularly for those with middle or low teleworking rates. However, high fluctuations in heart rate were a noticeable contributor to stress in those with a high teleworking rate (Figure 5C).

Figure 4. Analysis of the contribution direction of (A) working hours and (B) activity categorized by teleworking/telecommuting rates based on Shapley Additive Explanations (SHAP).

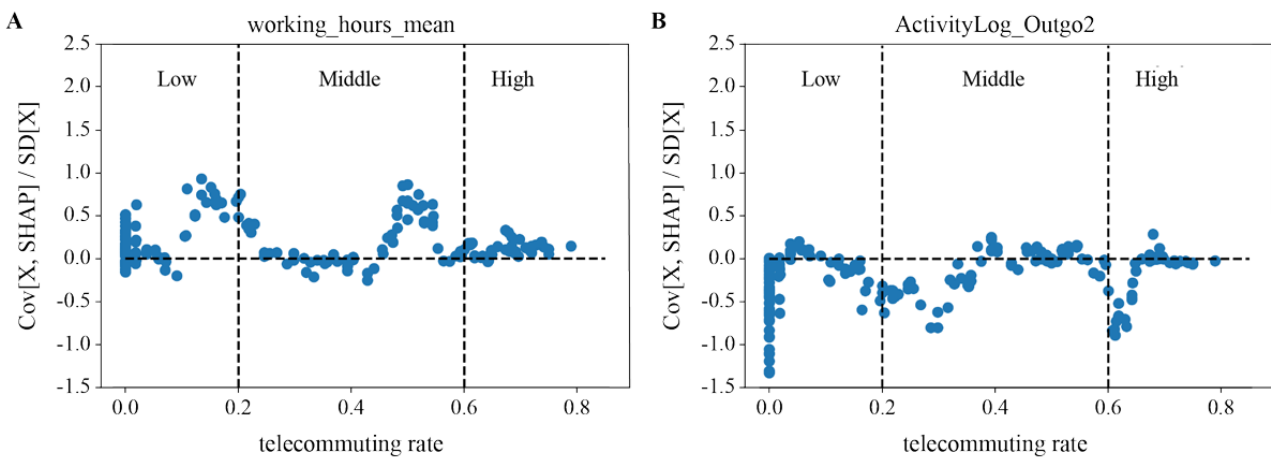
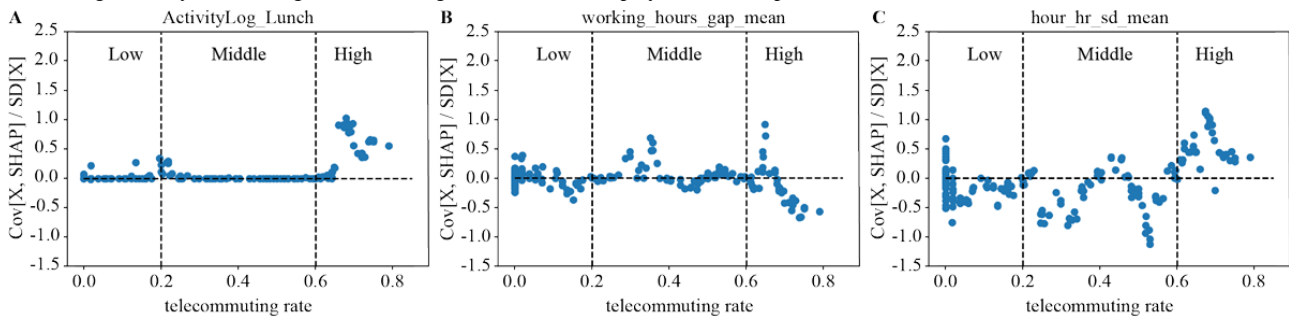


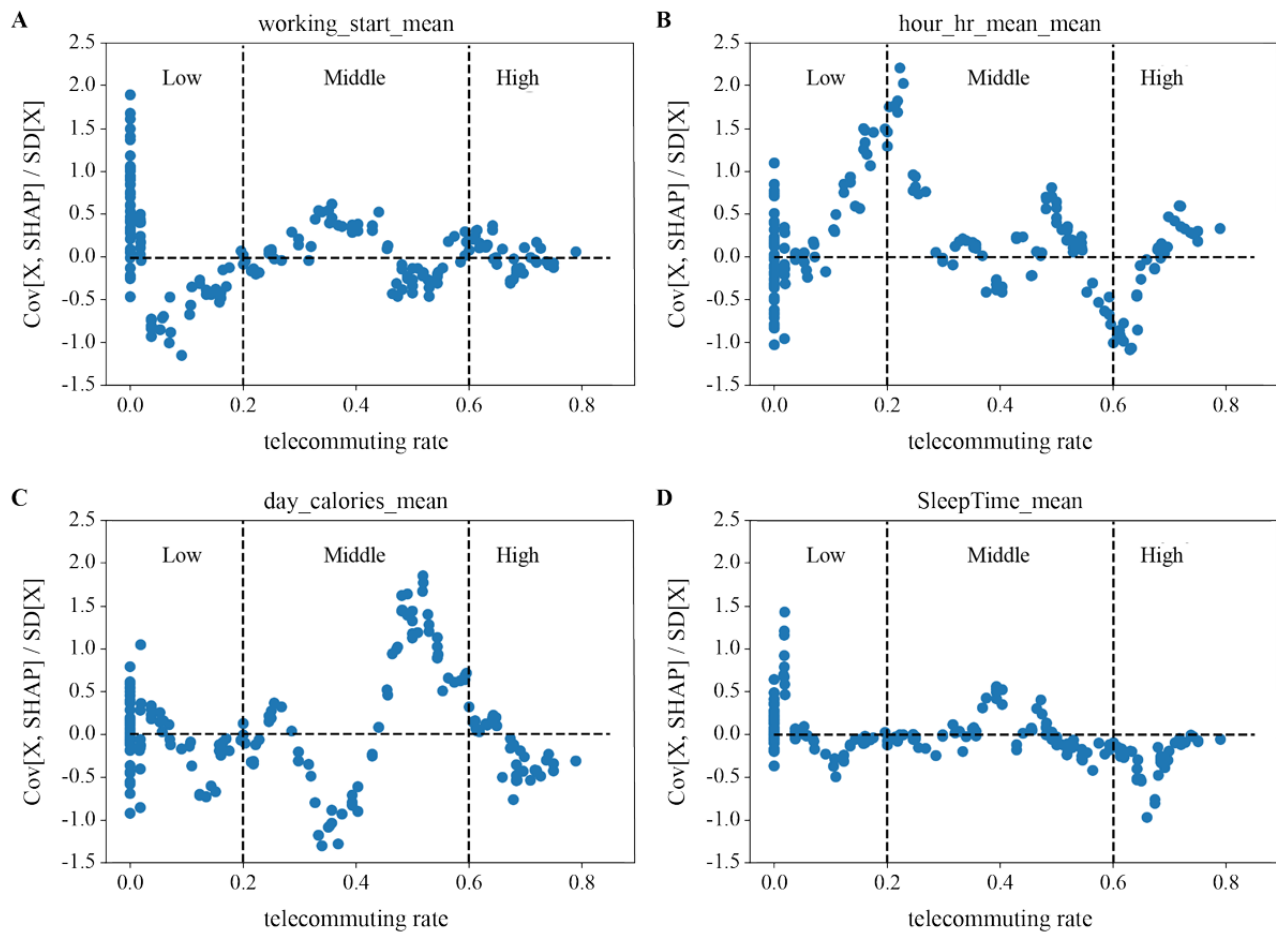
Figure 5. Analysis of the contribution direction of (A) skipping lunch, (B) working hour gap (working more or less than scheduled hours), and (C) heart rate categorized by teleworking/telecommuting rates based on Shapley Additive Explanations (SHAP).



In participants with low teleworking rates, being late for work or coming to work too early contributed to stress. Although the variation was lower, a similar trend was observed for participants with high and middle teleworking rates (Figure 6A). Having a heart rate higher or lower than the mean heart rate contributed to stress in participants with low teleworking rates. Although the variation was lower, a similar trend was observed for participants with high and middle teleworking rates (Figure 6B). Burning more or fewer calories than the mean

calorie burned contributed to stress in participants with middle and low teleworking rates. Moreover, burning less than normal calories was a noticeable contributor to stress in participants with high teleworking rates (Figure 6C). In participants with a low teleworking rate, a longer mean sleep duration contributed to stress, whereas in those with a high teleworking rate, a lower mean sleep duration was a noticeable contributor to stress (Figure 6D).

Figure 6. Analysis of the contribution direction of (A) mean work start time, (B) mean heart rate, (C) daily calories burned, and (D) sleep duration categorized by teleworking/telecommuting rates based on Shapley Additive Explanations (SHAP).



Discussion

Principal Findings

This study evaluated a novel, high-performance stress prediction algorithm that uses data from employees to extract neighborhood

data on working styles or work characteristics similar to those of the target person. The prediction performance of both proposed methods was markedly improved compared with that of the single model (baseline). A good stress prediction performance was achieved—the AUC was the highest for

proposed method 1 (0.85), followed by proposed method 2 (0.84) and the single model method (0.76). The level of predictive performance achieved by the proposed models suggested the benefits of narrowing the training data (by using neighborhood data) based on the teleworking rate.

In a stress detection study performed by Tazarv et al [30], per-individual models were reported to outperform single models; however, the approach required many data points (approximately 300 times/month) from participants. Therefore, by selecting a neighborhood cluster, the burden on participants was reduced. This approach alleviated user burden by reducing the number of label requests to 7 data points for the prediction target person. Because previous studies [20-22] did not narrow the training data based on work style/characteristics, it is possible to improve their prediction performance by incorporating this approach.

The results showed that personal data from the prediction target person are important (particularly in terms of measuring the change from baseline) because proposed method 2 showed prediction performance similar to that of proposed method 1. There was almost no difference in the AUC between proposed methods 1 and 2, suggesting that intraindividual fluctuation is a major stressor as the participants' own data contributed greatly to the performance prediction rather than the neighborhood cluster data. Thus, personal data from the prediction target person are important because a reduction in the neighborhood cluster's training data to 5 weeks caused no noticeable performance deterioration. Furthermore, the validity of using individual models is supported by the fact that there are differences in the feature contribution depending on the teleworking level, and the direction of the contribution changes within each level.

For participants with low teleworking rates, most features were related to sleep, followed by activity and work. This indicates that the contribution of activity may be lower when working from the office (low teleworking rates) than at other teleworking levels because it is difficult to discriminate between regular activity and activity due to commuting. For participants with high teleworking rates, most features were related to activity, followed by sleep and work. This implies that in a teleworking environment (such as at home), baseline activity levels are consciously assumed to be low and easier to discern than sleep and work.

The results of SHAP suggest that some features are consistent with intuition and common sense, contributing to its validity. Longer working hours among participants with middle and low teleworking rates were a marker of high stress. Low activity, irrespective of the number of days worked from the office per week, was a marker of high stress. Additionally, some features showed changes in the contribution direction within teleworking levels, suggesting the validity of the proposed method for modeling a small group of participants.

Several features characteristic of the high teleworking group, which tended to have the same working style among individuals but in a completely different working environment, were identified. Skipping lunch while working from home was likely to be a marker of stress. This could also be attributed to the fact

that with a high degree of freedom, a person is more likely to skip meals. In addition, biological information, such as skipping meals/hunger, is not as easily discernible by employees as activity, which is presumed to be low while teleworking. Additionally, working more or less than the scheduled hours contributed to stress, especially among those with a high teleworking rate. This observation suggested that arriving late or leaving early for appointments may be detected as a sign of stress, likely due to the high psychological hurdles for arriving late or leaving early, especially among those working from the office. We believe that psychological hurdles are fewer when working from home, possibly due to the higher degree of flexibility in using the provided working hours.

Additionally, lower fluctuations in heart rate were found to contribute to stress, especially in participants with middle and low teleworking rates. However, a higher fluctuation in heart rate was a noticeable contributor to stress in those with a high teleworking rate. Although it is known that the lower the fluctuations in heart rate, the greater the stress [9], contradictory results were noted in the high teleworking group. The autonomic nervous system, which consists of sympathetic and parasympathetic nerves, regulates heart rate. During a fight or flight response (work stress or activity in the contemporary sense), sympathetic nerves increase the heart rate. On the other hand, during the rest and digest state (relaxing or inactivity), the parasympathetic nerves dominate and decrease heart rate. It is assumed that sympathetic activation is dominant while working from the office and parasympathetic activation is dominant while teleworking [18]. The low fluctuations in heart rate associated with high stress levels in the low and middle teleworking groups could be attributed to sustained sympathetic dominance with less time to relax while working from the office. Similarly, high fluctuations in heart rate associated with high stress levels in the high teleworking group could be attributed to temporal activation of sympathetic nerves while performing a difficult task, despite the parasympathetic predominance of the baseline state. Additionally, a lower mean sleep duration among participants with a high teleworking rate was a marker of stress in this study. This result is important because we expect that a person should get sufficient sleep when working from home.

Similarly, several features characteristic of the low teleworking group were identified. Coming late or too early to work was identified as a marker of stress among those with a low teleworking rate. These observations suggested that coming too early may correlate with long working hours and coming late may correlate with decreased engagement. Moreover, having a higher or lower than mean heart rate was found to be a marker of stress in those with a low teleworking rate. This suggests that in terms of heart rate, an individual may respond differently to stress while working from the office, according to the baseline state of the autonomic nervous system with sympathetic or parasympathetic dominance. Moreover, the variability in the contribution of calories burned was high among those with middle and low teleworking rates. Burning more or fewer calories than normal among participants with middle and low teleworking rates was a marker of stress and could be attributed to the individual's unique response.

Limitations

The data used in this study (ie, wearable device, questionnaire, and attendance data) were affected by working style and various other factors. If the target population were to change, the results may be different from those obtained in this study. Moreover, age-related comorbidities and lifestyle changes were not considered in the modeling, which can impact the outcome. In this study, we created a neighborhood cluster based on the teleworking rate. Therefore, it can only be applied to people who are allowed to telework. The “neighborhood cluster” in this study was assumed to be a “cluster with similar working style.” For practical purposes, it is conceivable that working styles differ greatly, even if the teleworking rate is similar (eg, when data are obtained from multiple companies). Moreover, responses to the questionnaires, including the K6 questionnaire, were subjective for the participants and not necessarily accurate. Furthermore, feature importance and SHAP only quantify the degree to which the machine learning model uses the features for prediction but do not consider whether the model makes predictions with high accuracy. Thus, although the tendency to judge that stress is high when the value of a feature is large is correct, it cannot be confirmed that “stress increases when the

value of a feature is large.” Finally, because teleworking outside of working from home was not allowed in the Shionogi Group, a certain degree of participant bias may exist because certain job functions were not permitted to telework. Therefore, the results of this study might not be reproducible when targeting other forms of teleworking.

Conclusion

Prediction performance was improved by forming a cluster (neighborhood cluster) with similar working styles based on the teleworking rate and using it as the training data. The validity of the neighborhood cluster approach is indicated by differences in the contributing features and their contribution directions among teleworking levels. Further studies are required to evaluate and improve the proposed method using data obtained from employees of different companies. This methodology can improve existing stress detection methods by incorporating the idea of this research and narrowing the training data (ie, neighborhood cluster extraction based on the teleworking rate). This study paves the way for employers to consider and support timely and appropriate interventions for people predicted to experience high stress levels.

Acknowledgments

We thank Kazuhisa Nagaishi, Shogo Miyazawa, Yukichi Ishioka, Masahiko Oya, Yukiko Sawada, Aki Murakami, Yuichi Yamada, and Tomoko Yoshida of Shionogi & Co., Ltd for their contributions. Medical writing support was provided by Annirudha Chillar, MD, PhD, of Cactus Life Sciences (part of Cactus Communications) and was funded by Shionogi & Co., Ltd.

Conflicts of Interest

YA is a part-time employee as an industrial physician with Shionogi & Co., Ltd and has a patent issued (2023-062254). HI is a full-time employee of Shionogi & Co., Ltd and has received study funding from the company since the initial planning of the work. He also holds a patent (2023-062254). RK, YK, RT, YY, and SN are full-time employees of Shionogi & Co., Ltd and have received study funding from the company since the initial planning of the work. They also hold stocks via employee stock ownership society, along with a patent (2023-062254).

Multimedia Appendix 1

Variables evaluated to deduce the feature importance.

[[DOCX File, 30 KB - ai_v3i1e55840_app1.docx](#)]

Multimedia Appendix 2

Top 10 features with the highest mean feature importance ranking categorized into three levels of teleworking rates using Extreme Gradient Boosting (XGBoost). Features related to activity are in red, features related to work are in green, and features related to sleep are in blue.

[[DOCX File, 24 KB - ai_v3i1e55840_app2.docx](#)]

References

1. Yari beygi H, Panahi Y, Sahraei H, Johnston TP, Sahebkar A. The impact of stress on body function: A review. *EXCLI J* 2017;16:1057-1072 [[FREE Full text](#)] [doi: [10.17179/excli2017-480](https://doi.org/10.17179/excli2017-480)] [Medline: [28900385](https://pubmed.ncbi.nlm.nih.gov/28900385/)]
2. Hammen C. Stress and depression. *Annu Rev Clin Psychol* 2005;1:293-319. [doi: [10.1146/annurev.clinpsy.1.102803.143938](https://doi.org/10.1146/annurev.clinpsy.1.102803.143938)] [Medline: [17716090](https://pubmed.ncbi.nlm.nih.gov/17716090/)]
3. Towards a society where all people can play an active role in dealing with disabilities and illnesses. Ministry of Health, Labour and Welfare. URL: <https://www.mhlw.go.jp/stf/wp/hakusyo/kousei/18/index.html> [accessed 2023-12-12]
4. Burcusa SL, Iacono WG. Risk for recurrence in depression. *Clin Psychol Rev* 2007 Dec;27(8):959-985 [[FREE Full text](#)] [doi: [10.1016/j.cpr.2007.02.005](https://doi.org/10.1016/j.cpr.2007.02.005)] [Medline: [17448579](https://pubmed.ncbi.nlm.nih.gov/17448579/)]
5. Okumura Y, Higuchi T. Cost of depression among adults in Japan. *Prim Care Companion CNS Disord* 2011;13(3) [[FREE Full text](#)] [doi: [10.4088/PCC.10m01082](https://doi.org/10.4088/PCC.10m01082)] [Medline: [21977377](https://pubmed.ncbi.nlm.nih.gov/21977377/)]

6. Kraus C, Kadriu B, Lanzenberger R, Zarate CA, Kasper S. Prognosis and improved outcomes in major depression: a review. *Transl Psychiatry* 2019 Apr 03;9(1):127 [FREE Full text] [doi: [10.1038/s41398-019-0460-3](https://doi.org/10.1038/s41398-019-0460-3)] [Medline: [30944309](https://pubmed.ncbi.nlm.nih.gov/30944309/)]
7. 2021 Occupational Health and Safety Survey (Survey of Facts). Ministry of Health, Labour and Welfare. URL: <https://www.mhlw.go.jp/toukei/list/r03-46-50.html>; [accessed 2023-12-12]
8. Results of the 10th Corporate Questionnaire Survey on “Mental Health Initiatives”. Japan Productivity Division, Public Benefits Foundation. URL: <https://www.jpc-net.jp/research/detail/005595.html> [accessed 2023-12-12]
9. Kim H, Cheon E, Bai D, Lee YH, Koo B. Stress and heart rate variability: a meta-analysis and review of the literature. *Psychiatry Investig* 2018 Mar;15(3):235-245 [FREE Full text] [doi: [10.30773/pi.2017.08.17](https://doi.org/10.30773/pi.2017.08.17)] [Medline: [29486547](https://pubmed.ncbi.nlm.nih.gov/29486547/)]
10. Fox KR. The influence of physical activity on mental well-being. *Public Health Nutr* 1999 Sep;2(3A):411-418. [doi: [10.1017/s1368980099000567](https://doi.org/10.1017/s1368980099000567)] [Medline: [10610081](https://pubmed.ncbi.nlm.nih.gov/10610081/)]
11. Stein MB, Belik S, Jacobi F, Sareen J. Impairment associated with sleep problems in the community: relationship to physical and mental health comorbidity. *Psychosom Med* 2008 Oct;70(8):913-919. [doi: [10.1097/PSY.0b013e3181871405](https://doi.org/10.1097/PSY.0b013e3181871405)] [Medline: [18842741](https://pubmed.ncbi.nlm.nih.gov/18842741/)]
12. Stults-Kolehmainen MA, Sinha R. The effects of stress on physical activity and exercise. *Sports Med* 2014 Jan;44(1):81-121 [FREE Full text] [doi: [10.1007/s40279-013-0090-5](https://doi.org/10.1007/s40279-013-0090-5)] [Medline: [24030837](https://pubmed.ncbi.nlm.nih.gov/24030837/)]
13. Hakro S, Jameel A, Hussain A, Aslam MS, Khan WA, Sadiq S, et al. A lunch break time and its impact on employees health, performance and stress on work. *J Pharm Res Int* 2021 Jul 27:84-97. [doi: [10.9734/jpri/2021/v33i38b32102](https://doi.org/10.9734/jpri/2021/v33i38b32102)]
14. Tavares AI. Telework and health effects review. *Int J Healthc* 2017 Jul 11;3(2):30. [doi: [10.5430/ijh.v3n2p30](https://doi.org/10.5430/ijh.v3n2p30)]
15. Otsuka S, Ishimaru T, Nagata M, Tateishi S, Eguchi H, Tsuji M, CORoNaWork Project. A cross-sectional study of the mismatch between telecommuting preference and frequency associated with psychological distress among Japanese workers in the COVID-19 pandemic. *J Occup Environ Med* 2021 Sep 01;63(9):e636-e640. [doi: [10.1097/JOM.0000000000002318](https://doi.org/10.1097/JOM.0000000000002318)] [Medline: [34491971](https://pubmed.ncbi.nlm.nih.gov/34491971/)]
16. Okawara M, Yamashita S. A review of recent scientific findings on the health effects of working from home and implications for the development of regulations. *J Work Health Saf Regul* 2023 Oct 03;2023(1):1-16. [doi: [10.57523/jaohlev.ra.22-003](https://doi.org/10.57523/jaohlev.ra.22-003)]
17. Hall CE, Davidson L, Brooks SK, Greenberg N, Weston D. The relationship between homeworking during COVID-19 and both, mental health, and productivity: a systematic review. *BMC Psychol* 2023 Jun 27;11(1):188 [FREE Full text] [doi: [10.1186/s40359-023-01221-3](https://doi.org/10.1186/s40359-023-01221-3)] [Medline: [37370153](https://pubmed.ncbi.nlm.nih.gov/37370153/)]
18. Widar L, Wiitavaara B, Boman E, Heiden M. Psychophysiological reactivity, postures and movements among academic staff: a comparison between teleworking days and office days. *Int J Environ Res Public Health* 2021 Sep 10;18(18) [FREE Full text] [doi: [10.3390/ijerph18189537](https://doi.org/10.3390/ijerph18189537)] [Medline: [34574461](https://pubmed.ncbi.nlm.nih.gov/34574461/)]
19. Fukushima N, Machida M, Kikuchi H, Amagasa S, Hayashi T, Odagiri Y, et al. Associations of working from home with occupational physical activity and sedentary behavior under the COVID-19 pandemic. *J Occup Health* 2021 Jan;63(1):e12212 [FREE Full text] [doi: [10.1002/1348-9585.12212](https://doi.org/10.1002/1348-9585.12212)] [Medline: [33683779](https://pubmed.ncbi.nlm.nih.gov/33683779/)]
20. Saeed A, Trajanovski S. Personalized driver stress detection with multi-task neural networks using physiological signals. arXiv Preprint posted online on November 15, 2017. [doi: [10.48550/arXiv.1711.06116](https://doi.org/10.48550/arXiv.1711.06116)]
21. Can YS, Chalabianloo N, Ekiz D, Fernandez-Alvarez J, Riva G, Ersoy C. Personal stress-level clustering and decision-level smoothing to enhance the performance of ambulatory stress detection with smartwatches. *IEEE Access* 2020;8:38146-38163. [doi: [10.1109/access.2020.2975351](https://doi.org/10.1109/access.2020.2975351)]
22. Bin M, Khalifa O, Saeed R. Real-time personalized stress detection from physiological signals. 2016 Jan 14 Presented at: International Conference on Computing, Control, Networking, Electronics and Embedded Systems Engineering (ICCNEEE); 2015; Khartoum, Sudan p. 352-356. [doi: [10.1109/icneee.2015.7381390](https://doi.org/10.1109/icneee.2015.7381390)]
23. Sueki S. A review of the results of the Japanese version of K6 and results of a survey of Wako university newcomers. *Research Map*. 2020. URL: https://researchmap.jp/hajime_sueki/misc/30908071 [accessed 2024-07-07]
24. Kessler RC, Barker PR, Colpe LJ, Epstein JF, Gfroerer JC, Hiripi E, et al. Screening for serious mental illness in the general population. *Arch Gen Psychiatry* 2003 Feb;60(2):184-189. [doi: [10.1001/archpsyc.60.2.184](https://doi.org/10.1001/archpsyc.60.2.184)] [Medline: [12578436](https://pubmed.ncbi.nlm.nih.gov/12578436/)]
25. Veleva BI, van Bezooijen RL, Chel VGM, Numans ME, Caljouw MAA. Effect of ultraviolet light on mood, depressive disorders and well-being. *Photodermatol Photoimmunol Photomed* 2018 Sep;34(5):288-297. [doi: [10.1111/phpp.12396](https://doi.org/10.1111/phpp.12396)] [Medline: [29855075](https://pubmed.ncbi.nlm.nih.gov/29855075/)]
26. Luo C, Chen S, Chiang C, Wu W, Chen C, Chen W, et al. Association between ultraviolet b exposure levels and depression in Taiwanese adults: a nested case-control study. *Int J Environ Res Public Health* 2022 Jun 03;19(11) [FREE Full text] [doi: [10.3390/ijerph19116846](https://doi.org/10.3390/ijerph19116846)] [Medline: [35682430](https://pubmed.ncbi.nlm.nih.gov/35682430/)]
27. Friedman JH. Greedy function approximation: A gradient boosting machine. *Ann Statist* 2001 Oct 1;29(5):1189-1232. [doi: [10.1214/aos/1013203451](https://doi.org/10.1214/aos/1013203451)]
28. Chen T, Guestrin C. Xgboost: A scalable tree boosting system. 2016 Presented at: 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; August 13-17; San Francisco, CA. [doi: [10.1145/2939672.2939785](https://doi.org/10.1145/2939672.2939785)]
29. Lundberg S, Su-In LA. A unified approach to interpreting model predictions. 2017 Presented at: 31st Conference on Neural Information Processing Systems (NIPS –4777); December 5-7; Long Beach, CA p. 4768.

30. Tazarv A, Labbaf S, Reich S, Dutt N, Rahmani A, Levorato M. Personalized stress monitoring using wearable sensors in everyday settings. *Annu Int Conf IEEE Eng Med Biol Soc* 2021 Nov;2021:7332-7335. [doi: [10.1109/EMBC46164.2021.9630224](https://doi.org/10.1109/EMBC46164.2021.9630224)] [Medline: [34892791](https://pubmed.ncbi.nlm.nih.gov/34892791/)]

Abbreviations

AUROC: area under the receiver operating characteristic curve

FP: false positive

SHAP: Shapley Additive Explanations

TP: true positive

XGBoost: Extreme Gradient Boosting

Edited by K El Emam, B Malin; submitted 27.12.23; peer-reviewed by B Smarr, G Vos; comments to author 29.02.24; revised version received 18.04.24; accepted 14.06.24; published 02.08.24.

Please cite as:

Iwamoto H, Nakano S, Tajima R, Kiguchi R, Yoshida Y, Kitanishi Y, Aoki Y

Predicting Workers' Stress: Application of a High-Performance Algorithm Using Working-Style Characteristics

JMIR AI 2024;3:e55840

URL: <https://ai.jmir.org/2024/1/e55840>

doi: [10.2196/55840](https://doi.org/10.2196/55840)

PMID: [39093604](https://pubmed.ncbi.nlm.nih.gov/39093604/)

©Hiroki Iwamoto, Saki Nakano, Ryotaro Tajima, Ryo Kiguchi, Yuki Yoshida, Yoshitake Kitanishi, Yasunori Aoki. Originally published in JMIR AI (<https://ai.jmir.org>), 02.08.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR AI, is properly cited. The complete bibliographic information, a link to the original publication on <https://www.ai.jmir.org/>, as well as this copyright and license information must be included.

Original Paper

Optimizing Clinical Trial Eligibility Design Using Natural Language Processing Models and Real-World Data: Algorithm Development and Validation

Kyeryoung Lee^{1*}, PhD; Zongzhi Liu^{1*}, PhD; Yun Mai¹, PhD; Tomi Jun¹, MD; Meng Ma¹, PhD; Tongyu Wang¹, BSc; Lei Ai¹, PhD; Ediz Calay¹, PhD; William Oh^{1,2}, MD; Gustavo Stolovitzky¹, PhD; Eric Schadt^{1,2}, PhD; Xiaoyan Wang¹, PhD

¹GendDx (Sema4), Stamford, CT, United States

²Icahn School of Medicine at Mount Sinai, New York, NY, United States

*these authors contributed equally

Corresponding Author:

Xiaoyan Wang, PhD

GendDx (Sema4)

333 Ludlow Street

Stamford, CT, 06902

United States

Phone: 1 (844) 241 1233

Email: xw108@caa.columbia.edu

Abstract

Background: Clinical trials are vital for developing new therapies but can also delay drug development. Efficient trial data management, optimized trial protocol, and accurate patient identification are critical for reducing trial timelines. Natural language processing (NLP) has the potential to achieve these objectives.

Objective: This study aims to assess the feasibility of using data-driven approaches to optimize clinical trial protocol design and identify eligible patients. This involves creating a comprehensive eligibility criteria knowledge base integrated within electronic health records using deep learning-based NLP techniques.

Methods: We obtained data of 3281 industry-sponsored phase 2 or 3 interventional clinical trials recruiting patients with non-small cell lung cancer, prostate cancer, breast cancer, multiple myeloma, ulcerative colitis, and Crohn disease from ClinicalTrials.gov, spanning the period between 2013 and 2020. A customized bidirectional long short-term memory- and conditional random field-based NLP pipeline was used to extract all eligibility criteria attributes and convert hypernym concepts into computable hyponyms along with their corresponding values. To illustrate the simulation of clinical trial design for optimization purposes, we selected a subset of patients with non-small cell lung cancer (n=2775), curated from the Mount Sinai Health System, as a pilot study.

Results: We manually annotated the clinical trial eligibility corpus (485/3281, 14.78% trials) and constructed an eligibility criteria-specific ontology. Our customized NLP pipeline, developed based on the eligibility criteria-specific ontology that we created through manual annotation, achieved high precision (0.91, range 0.67-1.00) and recall (0.79, range 0.50-1) scores, as well as a high F_1 -score (0.83, range 0.67-1), enabling the efficient extraction of granular criteria entities and relevant attributes from 3281 clinical trials. A standardized eligibility criteria knowledge base, compatible with electronic health records, was developed by transforming hypernym concepts into machine-interpretable hyponyms along with their corresponding values. In addition, an interface prototype demonstrated the practicality of leveraging real-world data for optimizing clinical trial protocols and identifying eligible patients.

Conclusions: Our customized NLP pipeline successfully generated a standardized eligibility criteria knowledge base by transforming hypernym criteria into machine-readable hyponyms along with their corresponding values. A prototype interface integrating real-world patient information allows us to assess the impact of each eligibility criterion on the number of patients eligible for the trial. Leveraging NLP and real-world data in a data-driven approach holds promise for streamlining the overall clinical trial process, optimizing processes, and improving efficiency in patient identification.

KEYWORDS

natural language processing; real-world data; clinical trial eligibility criteria; eligibility criteria-specific ontology; clinical trial protocol optimization; data-driven approach

Introduction

Background

Clinical trials are crucial for developing new therapies, but they require significant resources and can introduce delays in drug development, leading to increased costs [1,2]. Complex and restrictive eligibility criteria hinder patient enrollment, impacting target goals, timelines, and ultimately patient well-being [3-5]. This issue is particularly notable in cancer trials with poor recruitment and high failure rates [6-8] because >80% of the trials fail to meet their initial target accruals and timelines [6,9]. In addition, overly restrictive eligibility criteria limit the representation of the broader patient population, reducing real-world applicability and treatment impact [10-13]. Nonetheless, the practice of trials reusing complicated eligibility criteria without a clear rationale is a common one [14], despite the minimal impact on trial outcomes [15]. Liu et al [15] demonstrated that broadening eligibility criteria using a data-driven approach can benefit initially excluded patients. A comprehensive and standardized eligibility criteria knowledge base that is compatible with real-world data can address these challenges. Such a knowledge base optimizes trial protocol design, improves patient enrollment, enhances the reliability and applicability of evidence synthesis, and fosters the efficient development of new therapies. Furthermore, it enables opportunities such as generating synthetic control arms (SCAs) for single-arm clinical trials using electronic health records (EHRs) [16-18].

The importance of semantically representing eligibility criteria interoperable with EHRs has been highlighted in multiple studies [19-21]. Converting free-text eligibility criteria to computable formats poses challenges, addressed by a range of natural language processing (NLP) techniques and transformer models [22-26]. An NLP interface, Criteria2Query, enables computable queries for eligible cohort identification using EHRs [27]. This tool supports clinical trial knowledge base development, enhancing EHR interoperability and scalability for efficient eligibility criteria knowledge engineering [28]. Manually annotated data sets such as “Chia, a large annotated corpus of clinical trial eligibility criteria” [29] and the “Leaf Clinical Trials corpus, the largest and most comprehensive human-annotated corpus of publicly available clinical trials eligibility criteria” [30] have significantly enhanced NLP model training and the development of effective query structures. Despite significant progress in bridging the gap between eligibility criteria and EHRs, limitations persist in accurately representing the granularities of eligibility criteria and real-time eligible patient number checks [20,31,32]. Using varying hierarchical levels of medical concepts, whether as hypernyms or hyponyms, presents one of the challenges when aligning eligibility criteria with EHRs; for instance, numerous trial eligibility criteria use hypernyms, which encompass a group of

related medical concepts, such as *cardiovascular disease*. Conversely, the patient problem list within the EHR specifies particular medical conditions or diseases (hyponyms), such as *myocardial infarction*. Establishing a standardized eligibility criteria knowledge base by transforming ambiguous hypernym concepts into computable hyponyms can enhance optimizing trial protocol design and identifying eligible patients through seamless integration with EHR data.

Objectives

In this study, we aim to create a standardized eligibility criteria knowledge base that seamlessly integrates with EHRs. By using deep learning-based NLP techniques, hypernym concepts in eligibility criteria will be converted to their EHR-compatible hyponyms along with their corresponding values. In addition, the prototype user interface will be developed as a pilot study, enabling the data-driven optimization of clinical trial protocols and the identification of eligible patients through the integration of the eligibility criteria knowledge base and EHRs.

Methods

Data Set

We obtained the data from ClinicalTrials.gov, specifically industry-sponsored phase 2 or 3 interventional clinical trials initiated between January 2013 and May 2020. A total of 3281 trials were identified: 817 (24.9%) for non-small cell lung cancer (NSCLC), 649 (19.78%) for prostate cancer (PCa), 1057 (32.22%) for breast cancer (BCa), 447 (13.62%) for multiple myeloma (MM), 160 (4.88%) for ulcerative colitis (UC), and 151 (4.6%) for Crohn disease (CD).

For the development of the prototype interface, we selected a subgroup of patients (n=2775) diagnosed with NSCLC from a previously curated cohort of patients with lung cancer. This cohort was established using the data from Mount Sinai-Sema4 Health System data [33], and patient information was deidentified for the purposes of this study.

Deep Learning-Based NLP Pipeline Development

Our NLP pipeline consists of 3 modules: ontology construction and manual annotation, model training and pipeline evaluation, and application.

Ontology Construction and Manual Annotation

To construct our ontology, we randomly selected 425 eligibility criteria from diverse cancer trials and manually analyzed entities and relations. This manual analysis focused on identifying entities and their relationships. Entities were subsequently categorized into primary and modifier groups, with detailed examples provided in [Multimedia Appendices 1 and 2](#). The primary groups included *demographic*, *diagnosis*, *biomarker*, *disease status*, *prior therapy*, *comorbidity*, *laboratory test*, *vital*, *procedure*, and *other medication*, while the modifier groups

included *value*, *condition*, *evidence*, *lines of therapy*, *negation*, *exception*, *grade*, *dose*, and *temporal*. Any entities that did not fall into the primary groups were classified as *other observation*. Furthermore, we defined relations between the entities. The commonly detected relationships between the *primary* and *modifier* groups were (1) *has_value_limit* between *demographic (age)* or *vital laboratory test* and *value*, (2) *has temporal limit* between *comorbidity* or *other medication* or *procedure* and *temporal*, (3) *has_negation* between *observation* or *biomarker* or *prior therapy* and *negation*, and (4) *has_exception* between *comorbidities* or *biomarker* or *diagnosis* and *exception*. Other relationships included *has_dose limit*, *has_line of therapy limit*, *has_grade limit*, *has_condition*, and *need_evidence*. The applicability of the ontology was tested on 60 UC and CD trials. Next, we manually annotated 246 eligibility criteria from NSCLC trials and performed model training using Clinical Language Annotation, Modeling, and Processing, which is an NLP toolkit [34].

Model Training and Pipeline Evaluation

A multilayer deep learning architecture was implemented for NLP modeling. The first step involved transforming the text into sequential vectors of characterization during the embedding process. These vectors were subsequently input into a bidirectional long short-term memory network, which is an artificial neural network designed for text classification. The bidirectional long short-term memory network was used to recognize patterns in both forward and backward directions [35]. The identified patterns were then passed to the next layer, which used a conditional random field model to compute the prediction probability [36]. The NLP model was trained using annotated criteria, with 80% of the manually annotated gold standard data allocated for training. Model performance was evaluated on a separate validation set (20%) using precision, recall, and F_1 -score values:

$$\text{Precision} = \text{TP} / (\text{TP} + \text{FP}) \quad (1)$$

$$\text{Recall} = \text{TP} / (\text{TP} + \text{FN}) \quad (2)$$

$$F_1\text{-score} = 2 \times (\text{Precision} \times \text{Recall}) / (\text{Precision} + \text{Recall}) \quad (3)$$

In equations 1 and 2, TP stands for true positives, FP for false positives, and FN for false negatives.

The manual annotation and training processes were iteratively performed with additional manually annotated notes until the model achieved a F_1 -score of >0.8 in the test set (Multimedia Appendix 3). To tailor the pipeline for specific cancer types, a preannotation method using the NSCLC pipeline was implemented for PCa, BCa, and MM for common eligibility criteria such as laboratory test values and comorbidities. Specific eligibility criteria such as biomarkers and treatments were manually annotated for each cancer type: PCa with 124 trials, BCa with 73 trials, and MM with 60 trials.

Application

The fully trained named entity recognition and relation models were integrated and applied to annotate the remaining eligibility criteria for the 4 types of cancer studied (BCa, MM, NSCLC, and PCa). The output data included sentences, tokens, parts of speech, entities, negations, and relations.

Construction of Standardized Eligibility Criteria Knowledge Base Table

The standardized knowledge base was constructed in an *EntityGroup-AttributeName-Value* format, involving 2 key steps: attribute normalization and transforming hypernyms to hyponyms with corresponding values.

Attribute Normalization

To normalize attributes, we used a 3-step approach. First, we assigned a Unified Medical Language System concept unique identifier to map synonyms of an entity, such as *estrogen receptor-positive*, *ER-positive*, and *ER+* to the Unified Medical Language System concept unique identifier C0279754. Second, we developed a set of rules (Table 1) to map abbreviations (eg, *CrCl* to *creatinine clearance*) and different phrases with the same meaning (eg, $\geq 1.5x$ ULN [where ULN stands for *upper limit of normal*], *greater than or equal to 1.5x ULN*, and $\geq 1.5x$ *upper limit of normal*) back to their original text. Finally, 2 domain experts manually curated unnormalized entities.

Table 1. Rules for attribute normalization.

Rule and attributes from eligibility criteria	Normalized attributes
Rules for mapping synonyms	
AST ^a , SGOT ^b , aspartate aminotransferase, serum AST	AST
ALT ^c , SGPT ^d , alanine aminotransferase, serum ALT	ALT
Total bilirubin, serum bilirubin, total bilirubin level, bilirubin level	Total bilirubin
Hgb ^e , hemoglobin	Hgb
HbA _{1c} ^f , hemoglobinA _{1c}	HbA _{1c}
serum creatinine, creatinine, creatinine levels, creatinine level	Serum creatinine
ANC ^g , absolute neutrophil count, absolute neutrophil counts, neutrophil count, neutrophil counts, absolute neutrophil	ANC
WBC ^h , white blood cells, white blood cell, WBC count, white blood cell count, white blood count, leucocytes	WBC
platelets, platelet, platelet count, platelet counts	Platelets
CrCl ⁱ , creatine clearance	CrCl
ALP ^j , alkaline phosphatase	ALP
ULN ^k , upper limit of normal	ULN
LLN ^l , lower limit of normal	LLN
Rules related to unit and temporal modifier	
less than or equal to, ≤	≤
greater than or equal to, ≥	≥
greater than, >	>
less than, <	<
within 4 weeks, within 28 days	within 4 weeks
within 2 weeks, within 14 days	within 2 weeks
within 3 weeks, within 21 days	within 3 weeks
last 6 months, past 6 months, within 6 months, within six months	within 6 months
last 3 months, past 3 months, within 3 months, within three months	within 3 months
within 2 years, last 2 years, past 2 years	within 2 years
within 3 years, last 3 years, past 3 years	within 3 years
within 5 years, last 5 years, past 5 years	within 5 years
10 ⁹ /L, 10 ⁹ /L, 10 ³ /uL, 10 ³ /microliter, 1000/uL, 1000/microliter, K/microliter, 10 ³ /mm ³	10 ³ /uL
Other miscellaneous rules	
Case insensitive	__m
Remove spaces	—

^aAST: aspartate aminotransferase.

^bSGOT: serum glutamic oxaloacetic transaminase.

^cALT: alanine transaminase.

^dSGPT: serum glutamic pyruvic transaminase.

^eHgb: hemoglobin.

^fHbA_{1c}: glycated hemoglobin.

^gANC: absolute neutrophil count.

^hWBC: white blood cell.

ⁱCrCl: creatinine clearance.

^jALP: alkaline phosphatase.

^kULN: upper limit of normal.

^lLLN: lower limit of normal.

^mNot applicable.

Transforming Hypernyms to Hyponyms Along With Corresponding Values

To formalize hypernyms, identified in primary groups such as *laboratory test*, *comorbidity*, *biomarker*, *prior therapy*, and *other medication*, we used the following approaches: (1) for *adequate organ function* *laboratory test* values, we determined prevalent laboratory test values by analyzing the unique laboratory test values for each test across the trials of the same cancer type that defined the normal organ function; and (2) for *comorbidity*, *biomarker*, *prior therapy*, and *other medication* hypernyms, we collected all example hyponyms described across the trials of the same cancer type.

Creation of a Prototype Interface for Enhancing Trial Protocol Design Optimization

We developed a prototype interface using the R programming language (R Foundation for Statistical Computing) and the *Shiny*

package to enhance trial protocol design optimization. The interface allows users to simulate the number of eligible patients based on specific criteria, including a combination of criteria such as histology, stages, laboratory test values, performance scores, prior line of therapy, and comorbidities. For this pilot study, a subset of patients with NSCLC (n=2775) was selected and deidentified. To ensure consistency and accuracy, we standardized the sample entities found in both the eligibility criteria knowledge base and EHRs using concept codes such as the *International Classification of Diseases*; Logical Observation Identifiers, Names, and Codes (LOINC); and normalized medical prescription codes. In addition, we converted the patients' absolute laboratory test values to either the upper limit of normal (ULN) or the lower limit of normal based on the provided normal ranges for each specific test. [Tables 2](#) and [3](#) and [Textbox 1](#) present some examples of normalized concepts and their codes.

Table 2. Examples of normalized codes for each concept and normal range of each laboratory test.

Laboratory test	LOINC ^a code	Normal range
ALT ^b (SGPT ^c ; U/L)	1742-6	7-56
AST ^d (SGOT ^e ; U/L)	1920-8	10-40
Total bilirubin in serum (mg/dL)	1975-2	0.1-1.2
Direct (conjugated) bilirubin in serum (mg/dL)	1968-7	<0.3
Serum creatinine (mg/dL)	2160-0	0.6-1.2 (male), 0.5-1.1 (female)
CrCl ^f (mL/min)	2164-2	97-137 (male), 88-128 (female)
ANC ^g (cells/ μ L)	26499-4	>90 mL/min/1.73 m ²
Platelets (cells/ μ L)	777-3	150,000-450,000
Hemoglobin (g/dL)	718-7	12-18

^aLOINC: Logical Observation Identifiers, Names, and Codes.

^bALT: alanine transaminase.

^cSGPT: serum glutamic pyruvic transaminase.

^dAST: aspartate aminotransferase.

^eSGOT: serum glutamic oxaloacetic transaminase.

^fCrCl: creatinine clearance.

^gANC: absolute neutrophil count.

Table 3. Examples of International Classification of Diseases, Tenth Revision (ICD-10), and International Classification of Diseases, Ninth Revision (ICD-9), disease codes.

Disease	ICD-10 codes	ICD-9 codes
Congestive heart failure	I50.2, I50.3, I50.4	428.[2-4][0-3]
Unstable angina	I20.0	411.1
Acute myocardial infarction	I21	410.9[0-2]
Arrhythmia	I49	429.9
Torsade de pointes	I45.81	426.82
Long QT syndrome	I45.81	426.82
Atrial fibrillation and flutter	I48	427.3[1-2]
Symptomatic bradycardia	R00.1	427.89
Uncontrolled hypertension	I10	401.[09]
Heart aneurysm	I25.3	414.1[09]
Coronary heart disease	I25.1	414.01
Cardiomyopathy	I42.9	425.[49]
Vasculitis, or angiitis	I77.6	447.6
Pericardial effusion	I31.3	423.9
Peripheral vascular disease	I73.9	443.9

Textbox 1. Examples of normalized medical prescription (RxNORM) drug codes.

Drug and RxNORM code
<ul style="list-style-type: none"> • Bortezomib: 356733 • Carfilzomib: 1302966 • Ixazomib: 1723735 • Lenalidomide: 342369 • Pomalidomide: 1369713

The interface uses a rule-based algorithm to match patients' EHR data with the criteria. The comprehensive rules for matching EHR data with criteria have been described in our previous studies [37]; for instance, we defined the following rules to map each laboratory test in EHRs to 1 corresponding LOINC code:

1. Mapping the laboratory test in the LOINC dictionary to the laboratory test in the EHR, based on the popularity rank available in the LOINC dictionary
2. Mapping the laboratory test for serum or plasma samples in the LOINC dictionary to the laboratory test in the EHR when the popularity rank is not available in the LOINC dictionary
3. If one-to-one mapping is not feasible using rule 1 and rule 2, the test unit (eg, *gram* is preferred *ovemolar*) is considered to facilitate the mapping
4. When one-to-one mapping is not attainable using rule 1, rule 2, and rule 3, preference is given to the laboratory test that lacks information about the method for mapping

We associated medication classes with their respective medications; for instance, we extended the annotation “post-menopausal not older than 60 years and taking LHRH

[luteinizing hormone–releasing hormone] agonist” to include “post-menopausal not older than 60 years and taking goserelin, leuprolide, or other LHRH agonists.” To achieve this, we used both our in-house knowledge bases and standard resources, such as the National Comprehensive Cancer Network's Clinical Practice Guidelines in Oncology.

Users can specify different criteria and combinations, such as different laboratory test values with specific conditions such as *no brain metastasis* to determine the number of qualified patients. The algorithm matches each patient's EHR data with the selected criteria and calculates the number of matched patients for each criterion. The performance of the interface was evaluated by comparing it to the manual patient selection process conducted by experienced clinical domain experts.

Ethical Considerations

This study was confirmed and approved by the Program for the Protection of Human Subjects at the Mount Sinai School of Medicine (IRB-17-01245)

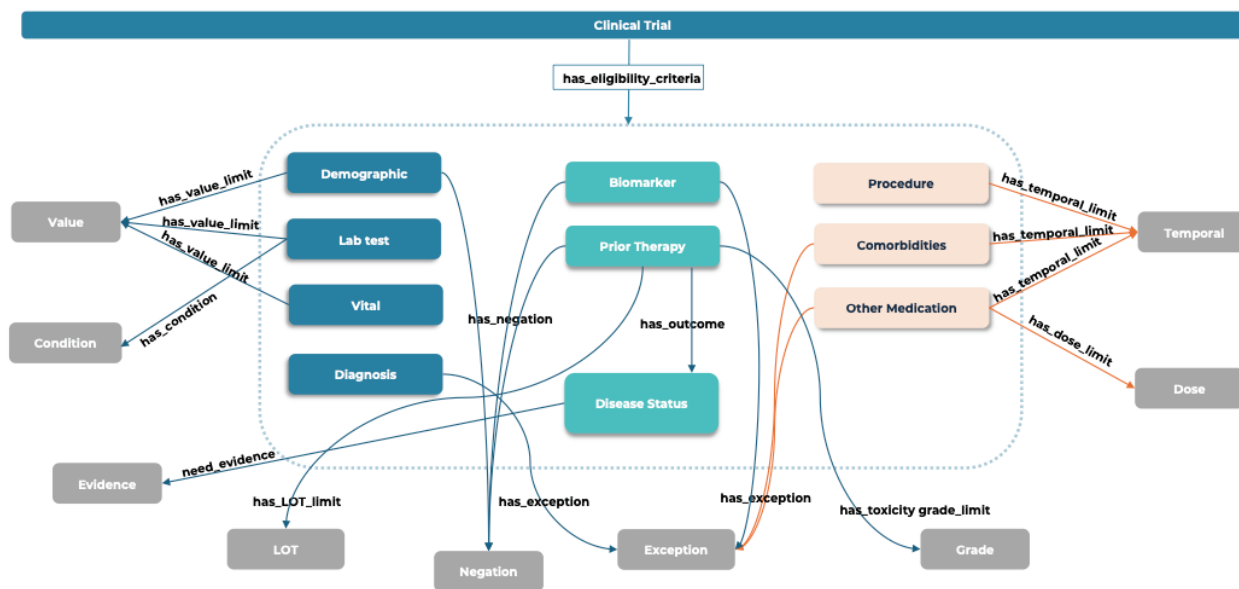
Results

Development of Eligibility Criteria–Specific Ontology

Our analysis of cancer clinical trials revealed that hormone therapy was the most frequently applied modality (1470/2970, 47.37%), primarily in BCa and PCa trials, followed by targeted therapy (753/2970, 25.35%) and immunotherapy (691/2970, 23.26%). Chemotherapy alone was used in 3.8% (113/2970) of the clinical trials. We developed an eligibility criteria ontology applicable to all cancer trials by manually analyzing 425 eligibility criteria (Figure 1). Entities were categorized into 10 primary groups (inside the blue dotted box) and 9 modifier groups based on semantic types and relations. Entities falling outside the blue dotted box were classified as *other observation*.

The inclusion criteria mainly involved entities in the *demographic, diagnosis, laboratory test, and vital* groups, while the exclusion criteria commonly included entities in the *comorbidity, procedure, and other medication* groups. Entities in the *biomarker, prior therapy, and disease status* groups appeared in both the inclusion and exclusion criteria. Relationships originated from the primary groups and terminated in the modifier groups, except for the *has outcome* relationship, which started and ended in the primary group (Figure 1). To assess the applicability of the cancer eligibility criteria ontology in a different disease context, we conducted a manual analysis of 60 trials related to UC and CD. For reference, the computable formats of the manually annotated 485 trials can be found in [Multimedia Appendices 4-8](#).

Figure 1. Clinical trial eligibility criteria ontology. Primary entities are grouped inside the blue dotted box. Modifier entities are placed outside the blue dotted box. The relationship between the primary entities and modifier entities always starts at a primary entity and ends at a modifier entity. LOT: line of therapy.



NLP Pipeline Quality Metrics

To evaluate the quality of our NLP pipeline, we computed precision, recall, and F_1 -score measures. For the primary group

entities, the average scores were 0.91 (precision), 0.79 (recall), and 0.83 (F_1 -score). Table 4 presents the range of precision, recall, and F_1 -score values of 17 primary group entities.

Table 4. Performance scores of customized natural language processing pipeline for each entity in the primary groups.

Primary group and attribute group::name	Precision	Recall	F ₁ -score
Demographic			
Demographic::age	1.000	0.923	0.960
Demographic::gender	1.000	0.870	0.931
Diagnosis			
Diagnosis::histology	1.000	1.000	1.000
Diagnosis::stage	0.667	1.000	0.800
Biomarker			
Biomarker::biomarker	1.000	0.800	0.889
Disease status			
Clinical status::disease status	0.737	0.684	0.709
Prior therapy			
Prior therapy::chemotherapy	0.944	0.895	0.919
Prior therapy::targeted therapy	1.000	0.786	0.880
Prior therapy::immunotherapy	0.897	0.788	0.839
Prior therapy::radiotherapy	0.682	0.682	0.682
Prior therapy::adjuvant therapy	1.000	0.571	0.727
Prior therapy::neoadjuvant therapy	1.000	0.500	0.667
Comorbidity			
Comorbidity::disease	0.842	0.762	0.800
Laboratory test			
Laboratory test::test	0.871	0.818	0.844
Vital			
Vital::vital	1.000	1.000	1.000
Procedure			
Procedure::procedure	1.000	0.600	0.750
Other medication			
Other medication::medication	0.800	0.727	0.762

Eligibility Criteria Attribute Extraction and Classification

The integrated named entity recognition and relation model extracted 9090 NSCLC, 7427 PCa, 10,217 BCa, 6803 MM, 1565 CD, and 1586 UC entities along with their attribute relations. After normalization and manual curation processes, the eligibility criteria knowledge base for each disease type was established in the *EntityGroup-AttributeName-Value* format ([Multimedia Appendices 9-14](#)). The number of unique *EntityGroup-AttributeName-Value* combinations varied across disease types, with 494 from 817 NSCLC trials, 471 from 649 PCa trials, 525 from 1057 BCa trials, 389 from 447 MM trials, 231 from 160 UC trials, and 230 from 151 CD trials. Notably, UC and CD trials had a smaller number of unique *EntityGroup-AttributeName-Value* combinations compared to cancer trials, indicating the presence of more complicated eligibility criteria in cancer trials.

[Figure 2](#) and [Table 5](#) show the distribution of *EntityGroup-AttributeName-Value* combinations in each primary group from different diseases and provide examples. The *laboratory test*, *prior therapy*, and *comorbidity* groups exhibited a high number of *EntityGroup-AttributeName-Value* combinations, followed by the *biomarker* and *other medication* groups. Variations were observed between solid cancers and hematologic cancers, with higher numbers of *EntityGroup-AttributeName-Value* combinations in solid cancer types for *prior therapy* and *biomarker*, while *laboratory test* and *comorbidity* were comparable. The *diagnosis* group exhibited varying numbers of *EntityGroup-AttributeName-Value* combinations across all 4 cancer types (BCa, MM, NSCLC, and PCa). *EntityGroup-AttributeName-Value* in the *biomarker*, *diagnosis*, and *prior therapy* groups were specified per indication, while shared *EntityGroup-AttributeName-Value* were found in other primary groups.

Figure 2. Distribution of attributes in the 10 primary groups as well as the other observation group extracted from the eligibility criteria of 4 different cancer types and 2 different autoimmune diseases. BCa: breast cancer; CD: Crohn disease; MM: multiple myeloma; NSCLC: non-small cell lung cancer; PCa: prostate cancer; UC: ulcerative colitis.

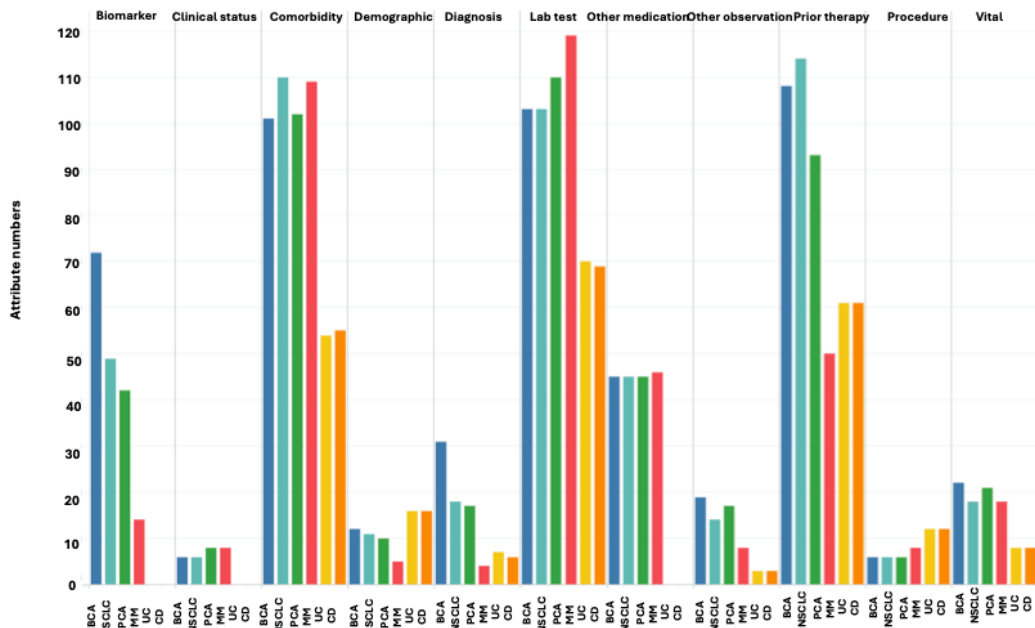


Table 5. The number of attributes for 10 primary groups along with examples.

Primary group	Number of attributes				Example attributes: group, name, value (with or without condition)
	NSCLC ^a	BCa ^b	PCa ^c	MM ^d	
Demographic	11	12	10	5	Demographic, age, ≥18 y
Diagnosis	18	31	17	4	Stage, TNM ^e system, T2b ^f
Biomarker	49	72	42	14	Biomarker, HER2 ^g mutation, L755P ^h
Disease status	11	11	13	9	Disease status, relapsed, yes
Prior therapy	114	108	93	50	LOT ⁱ , prior LOT, ≥2
Comorbidity	105	96	97	108	Cardiovascular disease, arrhythmia, yes (≤3 mo)
Laboratory test	103	103	110	119	Test, AST ^j , ≤2.5x ULN ^k
Vital	18	22	21	18	Vital, ECOG ^l , ≥2
Procedure	6	6	6	8	Procedure, organ transplantation, yes
Other medication	45	45	45	46	Other medication, use of anticoagulants, warfarin (<4 wk)

^aNSCLC: non-small cell lung cancer.

^bBCa: breast cancer.

^cPCa: prostate cancer.

^dMM: multiple myeloma.

^eTNM: tumor, nodes, metastasis.

^fT2b: a moderately advanced tumor in terms of size and extent but not the most advanced stage; specific implications can vary based on the type of cancer being described.

^gHER2: human epidermal growth factor receptor 2.

^hL755P: a reference to a specific mutation in the HER2 gene, with “L” standing for leucine, “755” being the position of the amino acid in the protein, and “P” standing for proline.

ⁱLOT: line of therapy.

^jAST: aspartate aminotransferase.

^kULN: upper limit of normal.

^lECOG: Eastern Cooperative Oncology Group.

Transformation of Umbrella Terms Into Computable Attributes With Representative Values

Overview

The conversion of hypernym concepts into computable attributes along with their corresponding values was carried out. [Table 6](#) provides some examples of converted attributes and their corresponding values for each hypernym. All lists can be found in [Multimedia Appendices 9-14](#).

Table 6. Examples of hypernym concepts (entity and subgroup entity in eligibility criteria) used in eligibility criteria and converted hyponyms along with their corresponding values.

Entity and subgroup entity in eligibility criteria and converted attribute	Corresponding values
Adequate organ function	
Normal hepatic function	
AST ^a	≤2.5x ULN ^b
ALT ^c	≤2.5x ULN
Total bilirubin	≤1.5x ULN
Normal renal function	
Creatinine	≤1.5x ULN
Normal hematologic function	
ANC ^d	≥1500 cells/uL
Platelets	≥100,000 cells/uL
Hemoglobin	≥9 mg/dL
Comorbidities	
Second malignancy	
All cancers	Yes, with exceptions
Infectious disease	
HIV	Yes
HBV ^e	Yes
HCV ^f	Yes
TB ^g	Yes
Cardiovascular disease	
CHF ^h	Yes
MI ⁱ	Yes
Angina	Yes
Arrhythmia	Yes
Autoimmune disease	
UC ^j	Yes
CD ^k	Yes
Systemic lupus erythematosus	Yes
Rheumatoid arthritis	Yes
Systemic sclerosis	Yes
Graves disease	Yes
Guillain-Barré syndrome	Yes
Antiphospholipid syndrome	Yes
Sjogren syndrome	Yes
Biomarker	
EGFR^l mutation sensitive to TKI^m	
Exon 19 deletion	Yes
Exon 21 L858R	Yes
Exon 21 L861Q	Yes

Entity and subgroup entity in eligibility criteria and converted attribute	Corresponding values
Exon 18 G719C	Yes
Exon 18 G719X	Yes
Amplification	Yes
EGFR mutation resistant to TKI	
Exon 20 T790M	Yes
Exon 20 C797S	Yes
Exon 20 S768I	Yes
Exon 20 insertion	Yes
Mismatch repair deficient	
MSH2, MSH6, MLH1, PMS2, or EXO1 gene mutation	Yes
MLH1 hypermethylation	Yes
Prior therapy (targeted)	
First-generation EGFR inhibitor	
Gefitinib	Yes
Erlotinib	Yes
Vandetanib	Yes
Second-generation EGFR inhibitor	
Afatinib	Yes
Dacomitinib	Yes
Poziotinib	Yes
Tesevatinib	Yes
Third-generation EGFR inhibitor	
Osimertinib	Yes
Lazertinib	Yes
Rociletinib	Yes
Tarloxotinib	Yes
Proteasome inhibitor	
Bortezomib based	Yes
Carfilzomib based	Yes
Ixazomib based	Yes
Oprozomib based	Yes
Prior therapy (hormone)	
First-generation antiandrogen	
Bicalutamide	Yes
Nilutamide	Yes
Flutamide	Yes
Second-generation antiandrogen	
Abiraterone	Yes
Enzalutamide	Yes
Darolutamide	Yes
Apalutamide	Yes
Androgen deprivation therapy	
Leuprolide	Yes

Entity and subgroup entity in eligibility criteria and converted attribute	Corresponding values
Goserelin	Yes
Degarelix	Yes
5-α reducing agent	
Finasteride	Yes
Dutasteride	Yes
Megestrol acetate	Yes
Other medication	
Current use of antibiotics	
Rifabutin	Yes
Clarithromycin	Yes
Azithromycin	Yes
Imipenem	Yes
Current use of antiarrhythmic agents	
Propafenone	Yes
Procainamide	Yes

^aAST: aspartate aminotransferase.

^bULN: upper limit of normal.

^cALT: alanine transaminase.

^dANC: absolute neutrophil count.

^eHBV: hepatitis B virus.

^fHCV: hepatitis C virus.

^gTB: tuberculosis.

^hCHF: congestive heart failure.

ⁱMI: myocardial infarction.

^jUC: ulcerative colitis.

^kCD: Crohn disease.

^lEGFR: epidermal growth factor receptor.

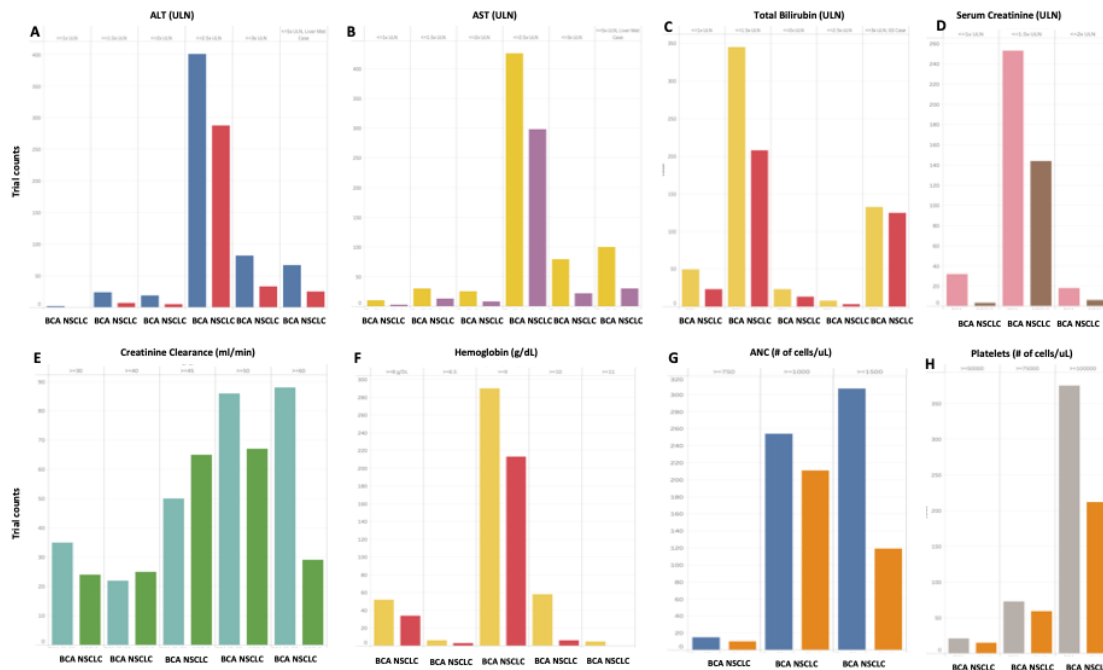
^mTKI: tyrosine kinase inhibitor.

Adequate Organ Function

Adequate organ function criteria were defined using various laboratory tests. Normal ranges and eligible values for alanine transaminase (ALT)/aspartate aminotransferase (AST), total bilirubin, serum creatinine, creatinine clearance, absolute neutrophil count, platelets, and hemoglobin were determined.

Representative values for *adequate organ/hematologic function* included $\leq 2.5x$ ULN for ALT/AST, $\leq 1.5x$ ULN for total bilirubin/serum creatinine, ≥ 1500 cells/uL for absolute neutrophil count, $\geq 100,000$ cells/uL for platelets, and ≥ 9 ng/dL for hemoglobin. [Figures 3A-3H](#) display the laboratory test value range and trial counts for each value in BCa and NSCLC clinical trials. The trends observed are similar in both cancer types.

Figure 3. Clinical trial counts with each unique laboratory test value defining normal organ function. (A-B) Alanine transaminase (ALT) and aspartate aminotransferase (AST): normal ranges from $\leq 1x$ upper limit of normal (ULN) to $\leq 3x$ ULN, with exceptions for liver diseases (eg, liver metastasis and Gilbert syndrome [GS]) allowing values of up to $\leq 5x$ ULN. (C) Total bilirubin: normal ranges from $\leq 1x$ ULN to $\leq 2.5x$ ULN, with exceptions for liver diseases (eg, liver metastasis and GS) allowing values of up to $\leq 3x$ ULN. (D) Serum creatinine: normal ranges from $\leq 1x$ ULN to $\leq 2.5x$ ULN. (E) Creatinine clearance: normal ranges from ≥ 30 to ≥ 60 mL/min. (F) Hemoglobin: normal ranges from ≥ 8.0 to ≥ 11.0 g/dL. (G) Absolute neutrophil count (ANC): normal ranges from ≥ 750 to ≥ 1500 cells/uL. (H) Platelets: normal ranges from $\geq 50,000$ to $\geq 100,000$ cells/uL. BCa: breast cancer; NSCLC: non-small cell lung cancer. For a higher-resolution version of this figure, see [Multimedia Appendix 15](#).

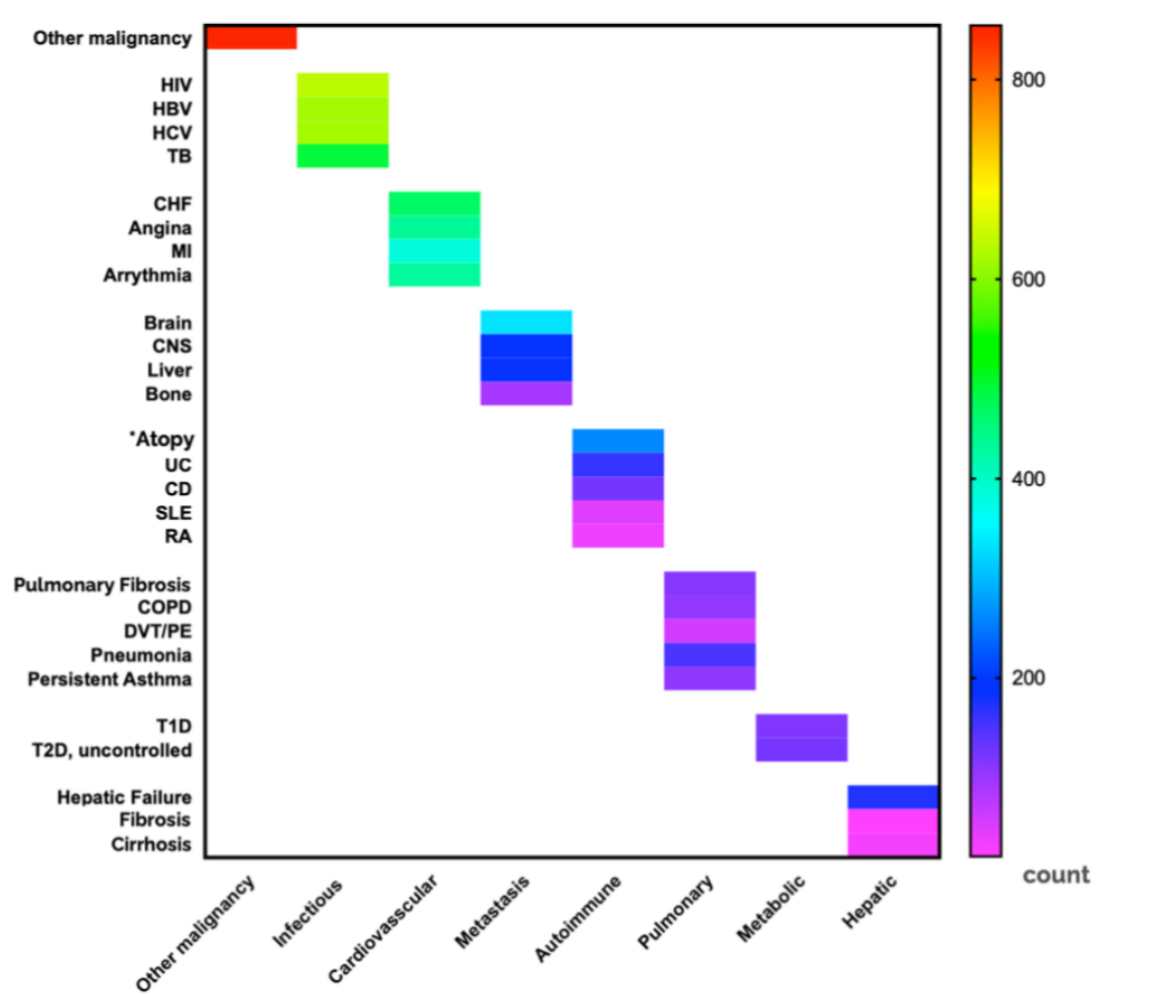


Comorbidities

The presence of comorbidities is a common exclusion criterion in clinical trials; however, natural language descriptions of comorbidities, such as “uncontrollable cardiovascular diseases,” “pulmonary diseases,” and “autoimmune diseases,” can be

ambiguous and need domain knowledge to interpret them. We analyzed the hypernyms and their corresponding hyponyms used in BCa trial eligibility criteria. [Figure 4](#) shows the collected hyponyms for each comorbidity class. The presence of second primary malignancies was excluded in almost all trials.

Figure 4. The heat map graph illustrates the number of clinical trials with each example hyponym for the hypernym comorbidities. Of note, the exception of atopy is mentioned as an autoimmune disease. The group does not include exceptions of other malignancies such as in situ cervical cancer, noninvasive bladder cancer, curative basal or squamous in situ prostate cancer, in situ breast cancer, or resected skin cancer other than melanoma. CD: Crohn disease; CHF: congestive heart failure; CNS: central nervous system; COPD: chronic obstructive pulmonary disease; DVT/PE: deep vein thrombosis/pulmonary embolism; HBV: hepatitis B virus; HCV: hepatitis C virus; MI: myocardial infarction; RA: rheumatoid arthritis; SLE: systemic lupus erythematosus; T1D: type 1 diabetes; T2D: type 2 diabetes; TB: tuberculosis; UC: ulcerative colitis.



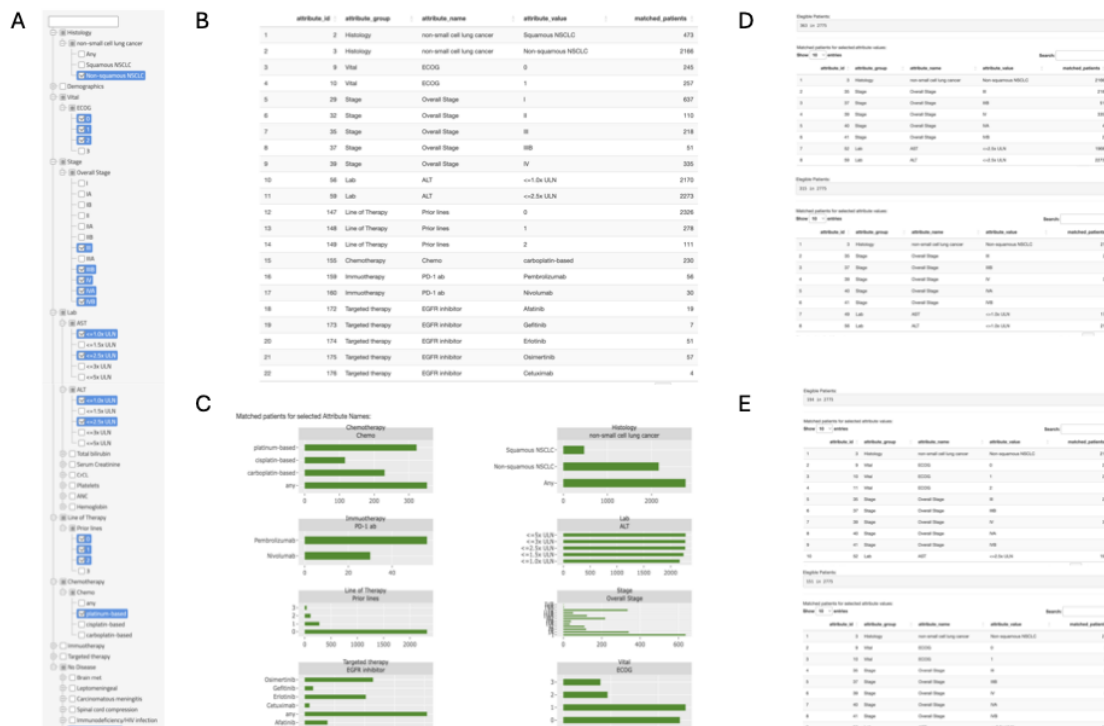
Prior Therapy, Other Medication, and Biomarker

By combining all examples of each hypernym, we broke down these hypernyms into actual medication and mutation hyponyms; for instance, we collected *procainamide* or *propafenone* for *current use of antiarrhythmic medication*. Similarly, we collected epidermal growth factor receptor (EGFR) exon 20 *T790M, T797S, S768I, or insertion* for *EGFR mutations resistant to EGFR inhibitors*.

Development of a Prototype Interface for the Optimization of Protocol Design

Our study investigated the impact of various criteria on the number of eligible patients. We developed a prototype interface that uses real-world patient information. Using a subset of deidentified cohorts of patients with NSCLC (n=2775), we deployed an eligibility criteria knowledge base that we had constructed in the interface. **Figure 5A** displays the selected criteria list, **Figure 5B** shows the corresponding patient number, and **Figure 5C** illustrates the distribution of patient numbers in each group.

Figure 5. Screenshots from a prototype interface. (A-B) The selected criteria list and the corresponding number of patients. (C) The distribution of patient numbers in each group. (D) Displayed are eligible patient numbers after sequentially incorporating criteria such as non-squamous histology and stage III and IV, with the further inclusion of aspartate aminotransferase (AST) and alanine transaminase (ALT) laboratory test values of either $\leq 2.5x$ upper limit of normal (ULN) or $\leq 1.0x$ ULN. (E) The influence of Eastern Cooperative Oncology Group (ECOG) performance status as an additional criterion. Displayed are eligible patient numbers after introducing ECOG scores of 0 to 2 or 0 to 1, with histology, stage, and ALT/AST laboratory test values ($< 2.5x$ ULN) as fixed criteria. ANC: absolute neutrophil count; CrCl: creatinine clearance; EGFR: epidermal growth factor receptor; NSCLC: non-small cell lung cancer; PD-1 ab: programmed cell death protein-1. For a higher-resolution version of this figure, see [Multimedia Appendix 16](#).



Sequentially incorporating criteria such as *nonsquamous histology* and *stages III and IV* criteria, we identified 2166 (78.05%) and 426 (15.35%) eligible patients, respectively, from the total pool of 2775 patients with NSCLC. The inclusion of *AST and ALT $\leq 2.5x$ ULN* criteria yielded 363 (13.08%) eligible patients from the pool of 2775 patients. Limiting AST and ALT to $\leq 1.0x$ ULN resulted in a decreased number of eligible patients (315/2775, 11.35%; [Figure 5D](#)). In addition, we explored the influence of Eastern Cooperative Oncology Group (ECOG) performance status as an additional criterion. With histology, stage, and ALT/AST laboratory test values ($< 2.5x$ ULN) as fixed criteria, by introducing ECOG scores of 0 to 2 or 0 to 1, we identified 194 (6.99%) and 151 (5.44%) eligible patients, respectively, from the pool of 2775 patients ([Figure 5E](#)).

Patient-matching performance was evaluated using precision, recall, and F_1 -score performance metrics across specific clinical attributes. The average F_1 -score, computed across 10 attributes from 8 domains (*other primary malignancy, congestive heart failure, squamous NSCLC, organ/tissue transplantation, platelets, programmed death-1 antibody therapy, programmed cell death protein-1 or programmed cell death program-ligand 1 positive, stage groups, prior LOT* [line of therapy], and *ECOG*), was 0.94 (range 0.82-1.00 [37]).

Discussion

Principal Findings

The challenge of achieving a high success rate in clinical trials is an ongoing issue [38,39]. Our study demonstrates the

feasibility of a data-driven approach to optimize trial protocols and efficiently identify eligible patients by constructing a comprehensive, EHR-interoperable eligibility criteria knowledge base and integrating EHR data. To accomplish this, we analyzed 3281 clinical trials using our customized deep learning NLP model. We extracted all entities with their attributes and converted the hypernym concepts used in eligibility criteria to EHR-compatible hyponyms along with their corresponding values. We also evaluated the feasibility of optimizing the trial protocol design on the interface we developed. This interface offers an efficient and effective approach for assessing the number of eligible patients across various combinations of eligibility criteria such as different laboratory test values as well as combinations that account for vital signs.

We developed an eligibility criteria-specific ontology by manually scrutinizing 425 eligibility criteria to be used as a reference for manual annotation during NLP model training. Accurately identifying intricate semantic relationships among entities within eligibility criteria is crucial for constructing an appropriate ontology for precise information extraction, including temporal, arithmetic values, Boolean values, and negation modifiers [31]. Our customized NLP pipeline based on the eligibility criteria-specific ontology that we created enabled us to efficiently extract all pertinent attributes across different modalities and diseases, allowing for a more accurate definition of the trial population. To determine the applicability of our ontology generated using cancer clinical trials to other disease domains, we compared the concepts and relations in clinical trials of inflammatory bowel diseases. We observed

very similar trends, suggesting that our eligibility criteria-specific ontology can be extended to other types of disease trials.

Moreover, the corpus of 485 manually annotated and standardized trials in a computable format can be used in eligible patient identification in EHRs.

Liu et al [28] conducted a thorough analysis of 352,100 clinical trials across various disease domains and constructed a knowledge base of clinical trial eligibility criteria. Their comprehensive knowledge base and user-friendly interface showcased the potential of advanced NLP techniques in enhancing eligibility criteria analysis and retrieval. Fang et al [40] also adopted a data-driven approach to optimizing clinical trial eligibility criteria in the context of Alzheimer disease and pancreatic cancer domains. Building upon these efforts, our study aimed to further narrow the gap between eligibility criteria and EHRs in multicancer domains, specifically in representing the granularities of eligibility criteria for identifying eligible patients and optimizing protocol designs. This was achieved by transforming hypernyms in the criteria into EHR-compatible hyponyms. We found that most of the primary groups include umbrella terms such as *prior therapy* (eg, proper prior therapy for actionable mutations) and *biomarker* (eg, EGFR inhibitor-resistant mutations). Our study also addressed the challenge of standardizing ambiguous clinical concepts in eligibility criteria for EHR interoperability and patient matching. To overcome this challenge, we converted hypernyms to the *Entity-Attribute-Value* format using prevailing values across different cancer types and modality therapies. We believe that our EHR-interoperable standardized eligibility criteria knowledge base and interface, integrating real-world EHR data, have the potential to improve the automatic screening system. This improvement has the potential to significantly reduce manual extraction efforts. Moreover, specific, computable criteria reduce ambiguity in patient identification and enable the inclusion of a broader range of patients who may qualify for the trial but could be excluded when using more general terms. This can increase patient trial enrollment, ultimately improving the overall success rate of trials. Notably, patients who were given the option to participate in a trial by their physicians demonstrated a significantly higher participation rate of 55% [41] compared to the current average of 5% to 8% among patients with cancer [42,43]. The implementation of our *hypernym/hyponym* semantic terminology model can likewise improve the effectiveness of information retrieval from EHRs and other clinical databases in the context of real-world evidence studies.

Certain criteria such as *histology*, *stage*, *previous treatment*, or *biomarker* are difficult to modify, while others such as vital signs or laboratory test values can be adjusted during the protocol design [15]. Our study revealed the impact of modifying laboratory test values while keeping other criteria constant, resulting in fluctuations in the number of eligible patients. Our findings, which demonstrate both the number of trials for different laboratory test value ranges and eligible patient numbers, offer insights for optimizing future protocol design and refining patient selection criteria. Seeking future collaboration with clinicians to conduct a direct comparison

between the patient identification results by clinical domain experts and those generated by our prototype holds promise for a more comprehensive and informative evaluation of the prototype's performance and its potential to enhance patient identification for clinical trials. Furthermore, a careful examination of the cases identified by the prototype can provide an understanding of the nature of false positives and false negatives. This will provide insights into how the prototype may differ in its patient identification results compared to manual extraction. Our eligibility criteria knowledge base can also be leveraged for generating SCAs using EHRs. SCAs, derived from real-world evidence, are regarded as substitutes for experimental control arms in trials [16-18]. The integration of SCAs into single-arm trial data or replacing traditional control arms with SCAs can alleviate the burden of target accrual in trials with low eligible patient numbers, such as rare disease or oncology trials with specific biomarkers. The Food and Drug Administration's approval of the palbociclib inhibitor for male patients with metastatic BCa based on real-world evidence demonstrates the potential and relevance of SCAs in improving trial design and outcomes [44].

Limitations

Our study has several limitations to consider. First, we focused on a limited scope, analyzing only 4 different cancer types and exploring extendibility in the context of inflammatory bowel diseases. Future studies should encompass a wider range of cancer types and disease domains for a more comprehensive analysis. Second, while most attributes were well defined, some umbrella terms lacked clear examples in other cancer types, potentially affecting result accuracy. Further manual annotation using knowledge bases could enhance the precision of the attribute tables. Third, our data set may be biased because we solely included industry-sponsored trials, potentially limiting the generalizability of our findings. In addition, the NLP training and test data sets in this study can display similarities owing to the shared attributes among different cancer trials, which heightens concerns regarding potential overfitting. Fourth, we did not address entity logic, and establishing the logic between entities would enhance cohort definition accuracy. Fifth and last, our interface feasibility testing was limited to small cohorts of patients with NSCLC, and the generalizability of our findings to other populations or disease conditions may vary. Furthermore, we did not perform a quantitative evaluation of the accuracy of matched patients although domain experts checked whether the patient information matched the eligibility criteria manually. While our model serves as a valuable illustration of how NLP can contribute to the design of trials across different diseases, we fully acknowledge the indispensable role of clinicians and biomedical researchers in ensuring the integrity of trial criteria. Clinical trials vary in their objectives, encompassing assessments of treatment end points, effectiveness, and other specific goals. The process is far more nuanced than merely adjusting laboratory test values because such modifications can have a substantial impact on the pool of eligible patients. Therefore, a comprehensive approach, considering both the clinical and biomedical aspects, is imperative for robust trial design.

Conclusions

Our study using an EHR-executable eligibility criteria knowledge base and real-world patient information provides

valuable insights into the influence of different criteria on the number of eligible patients during the protocol design. The findings highlight the potential of using a data-driven approach that incorporates NLP and EHRs in clinical research.

Data Availability

The patient data sets analyzed during this study are not publicly available due to patient privacy, security and the Health Insurance Portability and Accountability Act of 1996 (HIPAA) requirement. To enable a complete run of the code shared in this study, a minimum amount of desensitized sample data could be shared with the sharing agreement. Relevant requests should be addressed to ZL. The Prototype interface is available on the internet [45]; the R scripts are available on the internet [46].

Authors' Contributions

KL, ZL, and XW designed the study and wrote the manuscript. KL, YM, and XW reviewed the literature, clinical trial eligibility criteria, and patient notes and constructed the eligibility criteria-specific ontology. KL, ZL, MM, YM, and TW were responsible for the model training, the postprocessing procedure, and the data analysis. TJ, LA, EC, WO, GS, ES, and XW discussed the project and reviewed the manuscript.

Conflicts of Interest

KL, ZL, TJ, MM, YM, TW, LA, EC, GS, and XW are employees of Sema4. WO and ES are employees of the Icahn School of Medicine at Mount Sinai. WO holds stock in GeneDx. ES was employed by Sema4 and held stock in the company at the time the work in this paper was carried out.

Multimedia Appendix 1

Classification of attributes into primary entities and their subgroups.

[[PNG File , 79 KB - ai_v3i1e50800_app1.png](#)]

Multimedia Appendix 2

Classification of attributes into modifier entities and their subgroups.

[[PNG File , 62 KB - ai_v3i1e50800_app2.png](#)]

Multimedia Appendix 3

The scheme for the natural language processing-assisted clinical trial analysis.

[[PNG File , 50 KB - ai_v3i1e50800_app3.png](#)]

Multimedia Appendix 4

Manually annotated non-small cell lung cancer trials.

[[XLSX File \(Microsoft Excel File\), 184 KB - ai_v3i1e50800_app4.xlsx](#)]

Multimedia Appendix 5

Manually annotated breast cancer trials.

[[XLSX File \(Microsoft Excel File\), 133 KB - ai_v3i1e50800_app5.xlsx](#)]

Multimedia Appendix 6

Manually annotated prostate cancer trials.

[[XLSX File \(Microsoft Excel File\), 209 KB - ai_v3i1e50800_app6.xlsx](#)]

Multimedia Appendix 7

Manually annotated multiple myeloma trials.

[[XLSX File \(Microsoft Excel File\), 57 KB - ai_v3i1e50800_app7.xlsx](#)]

Multimedia Appendix 8

Manually annotated ulcerative colitis and Crohn disease trials.

[[XLSX File \(Microsoft Excel File\), 136 KB - ai_v3i1e50800_app8.xlsx](#)]

Multimedia Appendix 9

Non-small cell lung cancer attributes.

[[XLSX File \(Microsoft Excel File\), 28 KB - ai_v3ile50800_app9.xlsx](#)]

Multimedia Appendix 10

Breast cancer attributes.

[[XLSX File \(Microsoft Excel File\), 28 KB - ai_v3ile50800_app10.xlsx](#)]

Multimedia Appendix 11

Prostate cancer attributes.

[[XLSX File \(Microsoft Excel File\), 26 KB - ai_v3ile50800_app11.xlsx](#)]

Multimedia Appendix 12

Multiple myeloma attributes.

[[XLSX File \(Microsoft Excel File\), 23 KB - ai_v3ile50800_app12.xlsx](#)]

Multimedia Appendix 13

Ulcerative colitis attributes.

[[XLSX File \(Microsoft Excel File\), 29 KB - ai_v3ile50800_app13.xlsx](#)]

Multimedia Appendix 14

Crohn disease attributes.

[[XLSX File \(Microsoft Excel File\), 29 KB - ai_v3ile50800_app14.xlsx](#)]

Multimedia Appendix 15

Clinical trial counts with each unique laboratory test value defining normal organ function. (A-B) Alanine transaminase (ALT) and aspartate aminotransferase (AST): normal ranges from $\leq 1x$ upper limit of normal (ULN) to $\leq 3x$ ULN, with exceptions for liver diseases (eg, liver metastasis and Gilbert syndrome [GS]) allowing values of up to $\leq 5x$ ULN. (C) Total bilirubin: normal ranges from $\leq 1x$ ULN to $\leq 2.5x$ ULN, with exceptions for liver diseases (eg, liver metastasis and GS) allowing values of up to $\leq 3x$ ULN. (D) Serum creatinine: normal ranges from $\leq 1x$ ULN to $\leq 2.5x$ ULN. (E) Creatinine clearance: normal ranges from ≥ 30 to ≥ 60 mL/min. (F) Hemoglobin: normal ranges from ≥ 8.0 to ≥ 11.0 ng/dL. (G) Absolute neutrophil count (ANC): normal ranges from ≥ 750 to ≥ 1500 cells/uL. (H) Platelets: normal ranges from $\geq 50,000$ to $\geq 100,000$ cells/uL. BCa: breast cancer; NSCLC: non-small cell lung cancer.

[[PNG File , 122 KB - ai_v3ile50800_app15.png](#)]

Multimedia Appendix 16

Screenshots from a prototype interface. (A-B) The selected criteria list and the corresponding number of patients. (C) The distribution of patient numbers in each group. (D) Displayed are eligible patient numbers after sequentially incorporating criteria such as non-squamous histology and stage III and IV, with the further inclusion of aspartate aminotransferase (AST) and alanine transaminase (ALT) laboratory test values of either $\leq 2.5x$ upper limit of normal (ULN) or $\leq 1.0x$ ULN. (E) The influence of Eastern Cooperative Oncology Group (ECOG) performance status as an additional criterion. Displayed are eligible patient numbers after introducing ECOG scores of 0 to 2 or 0 to 1, with histology, stage, and ALT/AST laboratory test values ($< 2.5x$ ULN) as fixed criteria. ANC: absolute neutrophil count; CrCl: creatinine clearance; EGFR: epidermal growth factor receptor; NSCLC: non-small cell lung cancer; PD-1 ab: programmed cell death protein-1.

[[PNG File , 330 KB - ai_v3ile50800_app16.png](#)]

References

1. Dickson M, Gagnon JP. Key factors in the rising cost of new drug discovery and development. *Nat Rev Drug Discov* 2004 May;3(5):417-429. [doi: [10.1038/nrd1382](https://doi.org/10.1038/nrd1382)] [Medline: [15136789](https://pubmed.ncbi.nlm.nih.gov/15136789/)]
2. Wouters OJ, McKee M, Luyten J. Estimated research and development investment needed to bring a new medicine to market, 2009-2018. *JAMA* 2020 Mar 03;323(9):844-853 [FREE Full text] [doi: [10.1001/jama.2020.1166](https://doi.org/10.1001/jama.2020.1166)] [Medline: [32125404](https://pubmed.ncbi.nlm.nih.gov/32125404/)]
3. Jin S, Pazdur R, Sridhara R. Re-evaluating eligibility criteria for oncology clinical trials: analysis of investigational new drug applications in 2015. *J Clin Oncol* 2017 Nov 20;35(33):3745-3752 [FREE Full text] [doi: [10.1200/JCO.2017.73.4186](https://doi.org/10.1200/JCO.2017.73.4186)] [Medline: [28968168](https://pubmed.ncbi.nlm.nih.gov/28968168/)]

4. Garcia S, Bisen A, Yan J, Xie XJ, Ramalingam S, Schiller JH, et al. Thoracic oncology clinical trial eligibility criteria and requirements continue to increase in number and complexity. *J Thorac Oncol* 2017 Oct;12(10):1489-1495 [FREE Full text] [doi: [10.1016/j.jtho.2017.07.020](https://doi.org/10.1016/j.jtho.2017.07.020)] [Medline: [28802905](https://pubmed.ncbi.nlm.nih.gov/28802905/)]
5. Osarogiagbon RU, Vega DM, Fashoyin-Aje L, Wedam S, Ison G, Atienza S, et al. Modernizing clinical trial eligibility criteria: recommendations of the ASCO-friends of cancer research prior therapies work group. *Clin Cancer Res* 2021 May 01;27(9):2408-2415 [FREE Full text] [doi: [10.1158/1078-0432.CCR-20-3854](https://doi.org/10.1158/1078-0432.CCR-20-3854)] [Medline: [33563637](https://pubmed.ncbi.nlm.nih.gov/33563637/)]
6. Stensland KD, McBride RB, Latif A, Wisnivesky J, Hendricks R, Roper N, et al. Adult cancer clinical trials that fail to complete: an epidemic? *J Natl Cancer Inst* 2014 Sep;106(9):dju229. [doi: [10.1093/jnci/dju229](https://doi.org/10.1093/jnci/dju229)] [Medline: [25190726](https://pubmed.ncbi.nlm.nih.gov/25190726/)]
7. Fogel DB. Factors associated with clinical trials that fail and opportunities for improving the likelihood of success: a review. *Contemp Clin Trials Commun* 2018 Sep;11:156-164 [FREE Full text] [doi: [10.1016/j.conctc.2018.08.001](https://doi.org/10.1016/j.conctc.2018.08.001)] [Medline: [30112460](https://pubmed.ncbi.nlm.nih.gov/30112460/)]
8. Nipp RD, Hong K, Paskett ED. Overcoming barriers to clinical trial enrollment. *Am Soc Clin Oncol Educ Book* 2019 May;39:105-114. [doi: [10.1200/edbk.243729](https://doi.org/10.1200/edbk.243729)]
9. Brøgger-Mikkelsen M, Ali Z, Zibert JR, Andersen AD, Thomsen SF. Online patient recruitment in clinical trials: systematic review and meta-analysis. *J Med Internet Res* 2020 Nov 04;22(11):e22179 [FREE Full text] [doi: [10.2196/22179](https://doi.org/10.2196/22179)] [Medline: [33146627](https://pubmed.ncbi.nlm.nih.gov/33146627/)]
10. Naci H, Ioannidis JP. How good is "evidence" from clinical studies of drug effects and why might such evidence fail in the prediction of the clinical utility of drugs? *Annu Rev Pharmacol Toxicol* 2015;55:169-189. [doi: [10.1146/annurev-pharmtox-010814-124614](https://doi.org/10.1146/annurev-pharmtox-010814-124614)] [Medline: [25149917](https://pubmed.ncbi.nlm.nih.gov/25149917/)]
11. Heneghan C, Goldacre B, Mahtani KR. Why clinical trial outcomes fail to translate into benefits for patients. *Trials* 2017 Mar 14;18(1):122 [FREE Full text] [doi: [10.1186/s13063-017-1870-2](https://doi.org/10.1186/s13063-017-1870-2)] [Medline: [28288676](https://pubmed.ncbi.nlm.nih.gov/28288676/)]
12. Van Spall HG, Toren A, Kiss A, Fowler RA. Eligibility criteria of randomized controlled trials published in high-impact general medical journals: a systematic sampling review. *JAMA* 2007 Mar 21;297(11):1233-1240. [doi: [10.1001/jama.297.11.1233](https://doi.org/10.1001/jama.297.11.1233)] [Medline: [17374817](https://pubmed.ncbi.nlm.nih.gov/17374817/)]
13. Sen A, Ryan PB, Goldstein A, Chakrabarti S, Wang S, Koski E, et al. Correlating eligibility criteria generalizability and adverse events using Big Data for patients and clinical trials. *Ann N Y Acad Sci* 2017 Jan;1387(1):34-43 [FREE Full text] [doi: [10.1111/nyas.13195](https://doi.org/10.1111/nyas.13195)] [Medline: [27598694](https://pubmed.ncbi.nlm.nih.gov/27598694/)]
14. Begg CB. Cancer clinical trials in the USA: patient eligibility, generalizability of results and technology transfer. *Bull Cancer* 1987;74(2):197-203. [Medline: [3607303](https://pubmed.ncbi.nlm.nih.gov/3607303/)]
15. Liu R, Rizzo S, Whipple S, Pal N, Pineda AL, Lu M, et al. Evaluating eligibility criteria of oncology trials using real-world data and AI. *Nature* 2021 Apr;592(7855):629-633 [FREE Full text] [doi: [10.1038/s41586-021-03430-5](https://doi.org/10.1038/s41586-021-03430-5)] [Medline: [33828294](https://pubmed.ncbi.nlm.nih.gov/33828294/)]
16. Ko YA, Chen Z, Liu C, Hu Y, Quyyumi AA, Waller LA, et al. Developing a synthetic control group using electronic health records: application to a single-arm lifestyle intervention study. *Prev Med Rep* 2021 Dec;24:101572. [doi: [10.1016/j.pmedr.2021.101572](https://doi.org/10.1016/j.pmedr.2021.101572)] [Medline: [34976636](https://pubmed.ncbi.nlm.nih.gov/34976636/)]
17. Carrigan G, Whipple S, Capra WB, Taylor MD, Brown JS, Lu M, et al. Using electronic health records to derive control arms for early phase single-arm lung cancer trials: proof-of-concept in randomized controlled trials. *Clin Pharmacol Ther* 2020 Feb;107(2):369-377 [FREE Full text] [doi: [10.1002/cpt.1586](https://doi.org/10.1002/cpt.1586)] [Medline: [31350853](https://pubmed.ncbi.nlm.nih.gov/31350853/)]
18. Thomas DS, Lee AY, Müller PL, Schwartz R, Olvera-Barrios A, Warwick AN, et al. Contextualizing single-arm trials with real-world data: an emulated target trial comparing therapies for neovascular age-related macular degeneration. *Clin Transl Sci* 2021 May;14(3):1166-1175 [FREE Full text] [doi: [10.1111/cts.12974](https://doi.org/10.1111/cts.12974)] [Medline: [33421321](https://pubmed.ncbi.nlm.nih.gov/33421321/)]
19. Weng C, Tu SW, Sim I, Richesson R. Formal representation of eligibility criteria: a literature review. *J Biomed Inform* 2010 Jun;43(3):451-467 [FREE Full text] [doi: [10.1016/j.jbi.2009.12.004](https://doi.org/10.1016/j.jbi.2009.12.004)] [Medline: [20034594](https://pubmed.ncbi.nlm.nih.gov/20034594/)]
20. Chondrogiannis E, Andronikou V, Tagaris A, Karanastasis E, Varvarigou T, Tsuji M. A novel semantic representation for eligibility criteria in clinical trials. *J Biomed Inform* 2017 May;69:10-23 [FREE Full text] [doi: [10.1016/j.jbi.2017.03.013](https://doi.org/10.1016/j.jbi.2017.03.013)] [Medline: [28336477](https://pubmed.ncbi.nlm.nih.gov/28336477/)]
21. Su Q, Cheng G, Huang J. A review of research on eligibility criteria for clinical trials. *Clin Exp Med* 2023 Oct;23(6):1867-1879 [FREE Full text] [doi: [10.1007/s10238-022-00975-1](https://doi.org/10.1007/s10238-022-00975-1)] [Medline: [36602707](https://pubmed.ncbi.nlm.nih.gov/36602707/)]
22. Tu SW, Peleg M, Carini S, Bobak M, Ross J, Rubin D, et al. A practical method for transforming free-text eligibility criteria into computable criteria. *J Biomed Inform* 2011 Apr;44(2):239-250 [FREE Full text] [doi: [10.1016/j.jbi.2010.09.007](https://doi.org/10.1016/j.jbi.2010.09.007)] [Medline: [20851207](https://pubmed.ncbi.nlm.nih.gov/20851207/)]
23. Weng C, Wu X, Luo Z, Boland MR, Theodoratos D, Johnson SB. EliXR: an approach to eligibility criteria extraction and representation. *J Am Med Inform Assoc* 2011 Dec;18 Suppl 1(Suppl 1):i116-i124 [FREE Full text] [doi: [10.1136/amiajnl-2011-000321](https://doi.org/10.1136/amiajnl-2011-000321)] [Medline: [21807647](https://pubmed.ncbi.nlm.nih.gov/21807647/)]
24. Kang T, Zhang S, Tang Y, Hrubby GW, Rusanov A, Elhadad N, et al. EliIE: an open-source information extraction system for clinical trial eligibility criteria. *J Am Med Inform Assoc* 2017 Nov 01;24(6):1062-1071 [FREE Full text] [doi: [10.1093/jamia/ocx019](https://doi.org/10.1093/jamia/ocx019)] [Medline: [28379377](https://pubmed.ncbi.nlm.nih.gov/28379377/)]
25. Tian S, Erdengasileng A, Yang X, Guo Y, Wu Y, Zhang J, et al. Transformer-based named entity recognition for parsing clinical trial eligibility criteria. *ACM BCB* 2021 Aug;2021:49 [FREE Full text] [doi: [10.1145/3459930.3469560](https://doi.org/10.1145/3459930.3469560)] [Medline: [34414397](https://pubmed.ncbi.nlm.nih.gov/34414397/)]

26. Kreimeyer K, Foster M, Pandey A, Arya N, Halford G, Jones SF, et al. Natural language processing systems for capturing and standardizing unstructured clinical information: a systematic review. *J Biomed Inform* 2017 Sep;73:14-29 [FREE Full text] [doi: [10.1016/j.jbi.2017.07.012](https://doi.org/10.1016/j.jbi.2017.07.012)] [Medline: [28729030](https://pubmed.ncbi.nlm.nih.gov/28729030/)]
27. Yuan C, Ryan PB, Ta C, Guo Y, Li Z, Hardin J, et al. Criteria2Query: a natural language interface to clinical databases for cohort definition. *J Am Med Inform Assoc* 2019 Apr 01;26(4):294-305 [FREE Full text] [doi: [10.1093/jamia/ocy178](https://doi.org/10.1093/jamia/ocy178)] [Medline: [30753493](https://pubmed.ncbi.nlm.nih.gov/30753493/)]
28. Liu H, Chi Y, Butler A, Sun Y, Weng C. A knowledge base of clinical trial eligibility criteria. *J Biomed Inform* 2021 May;117:103771 [FREE Full text] [doi: [10.1016/j.jbi.2021.103771](https://doi.org/10.1016/j.jbi.2021.103771)] [Medline: [33813032](https://pubmed.ncbi.nlm.nih.gov/33813032/)]
29. Kury F, Butler A, Yuan C, Fu LH, Sun Y, Liu H, et al. Chia, a large annotated corpus of clinical trial eligibility criteria. *Sci Data* 2020 Aug 27;7(1):281 [FREE Full text] [doi: [10.1038/s41597-020-00620-0](https://doi.org/10.1038/s41597-020-00620-0)] [Medline: [32855408](https://pubmed.ncbi.nlm.nih.gov/32855408/)]
30. Dobbins NJ, Mullen T, Uzuner Ö, Yetisgen M. The Leaf Clinical Trials Corpus: a new resource for query generation from clinical trial eligibility criteria. *Sci Data* 2022 Aug 11;9(1):490 [FREE Full text] [doi: [10.1038/s41597-022-01521-0](https://doi.org/10.1038/s41597-022-01521-0)] [Medline: [35953524](https://pubmed.ncbi.nlm.nih.gov/35953524/)]
31. Ross J, Tu S, Carini S, Sim I. Analysis of eligibility criteria complexity in clinical trials. *Summit Transl Bioinform* 2010 Mar 01;2010:46-50 [FREE Full text] [Medline: [21347148](https://pubmed.ncbi.nlm.nih.gov/21347148/)]
32. Sim I, Olasov B, Carini S. An ontology of randomized controlled trials for evidence-based practice: content specification and evaluation using the competency decomposition method. *J Biomed Inform* 2004 Apr;37(2):108-119 [FREE Full text] [doi: [10.1016/j.jbi.2004.03.001](https://doi.org/10.1016/j.jbi.2004.03.001)] [Medline: [15120657](https://pubmed.ncbi.nlm.nih.gov/15120657/)]
33. Lee K, Liu Z, Chandran U, Kalsekar I, Laxmanan B, Higashi MK, et al. Detecting ground glass opacity features in patients with lung cancer: automated extraction and longitudinal analysis via deep learning-based natural language processing. *JMIR AI* 2023 Jun 1;2:e44537. [doi: [10.2196/44537](https://doi.org/10.2196/44537)]
34. Soysal E, Wang J, Jiang M, Wu Y, Pakhomov S, Liu H, et al. CLAMP - a toolkit for efficiently building customized clinical natural language processing pipelines. *J Am Med Inform Assoc* 2018 Mar 01;25(3):331-336 [FREE Full text] [doi: [10.1093/jamia/ocx132](https://doi.org/10.1093/jamia/ocx132)] [Medline: [29186491](https://pubmed.ncbi.nlm.nih.gov/29186491/)]
35. Alfattni G, Peek N, Nenadic G. Attention-based bidirectional long short-term memory networks for extracting temporal relationships from clinical discharge summaries. *J Biomed Inform* 2021 Nov;123:103915 [FREE Full text] [doi: [10.1016/j.jbi.2021.103915](https://doi.org/10.1016/j.jbi.2021.103915)] [Medline: [34600144](https://pubmed.ncbi.nlm.nih.gov/34600144/)]
36. Xu J, Li Z, Wei Q, Wu Y, Xiang Y, Lee HJ, et al. Applying a deep learning-based sequence labeling approach to detect attributes of medical concepts in clinical text. *BMC Med Inform Decis Mak* 2019 Dec 05;19(Suppl 5):236 [FREE Full text] [doi: [10.1186/s12911-019-0937-2](https://doi.org/10.1186/s12911-019-0937-2)] [Medline: [31801529](https://pubmed.ncbi.nlm.nih.gov/31801529/)]
37. Lee K, Mai Y, Liu Z, Raja K, Higashi MK, Jun T, et al. Establishing the automatic identification of clinical trial cohorts from electronic health records by matching normalized eligibility criteria and patient clinical characteristics. medRxiv Preprint posted online April 09, 2024 [FREE Full text] [doi: [10.1101/2024.02.28.24303396](https://doi.org/10.1101/2024.02.28.24303396)]
38. Hay M, Thomas DW, Craighead JL, Economides C, Rosenthal J. Clinical development success rates for investigational drugs. *Nat Biotechnol* 2014 Jan;32(1):40-51. [doi: [10.1038/nbt.2786](https://doi.org/10.1038/nbt.2786)] [Medline: [24406927](https://pubmed.ncbi.nlm.nih.gov/24406927/)]
39. Wong CH, Siah KW, Lo AW. Estimation of clinical trial success rates and related parameters. *Biostatistics* 2019 Apr 01;20(2):273-286 [FREE Full text] [doi: [10.1093/biostatistics/kxx069](https://doi.org/10.1093/biostatistics/kxx069)] [Medline: [29394327](https://pubmed.ncbi.nlm.nih.gov/29394327/)]
40. Fang Y, Liu H, Idnay B, Ta C, Marder K, Weng C. A data-driven approach to optimizing clinical study eligibility criteria. *J Biomed Inform* 2023 Jun;142:104375. [doi: [10.1016/j.jbi.2023.104375](https://doi.org/10.1016/j.jbi.2023.104375)] [Medline: [37141977](https://pubmed.ncbi.nlm.nih.gov/37141977/)]
41. Unger JM, Hershman DL, Till C, Minasian LM, Osarogiagbon RU, Fleury ME, et al. "When offered to participate": a systematic review and meta-analysis of patient agreement to participate in cancer clinical trials. *J Natl Cancer Inst* 2021 Mar 01;113(3):244-257 [FREE Full text] [doi: [10.1093/jnci/djaa155](https://doi.org/10.1093/jnci/djaa155)] [Medline: [33022716](https://pubmed.ncbi.nlm.nih.gov/33022716/)]
42. Unger JM, Vaidya R, Hershman DL, Minasian LM, Fleury ME. Systematic review and meta-analysis of the magnitude of structural, clinical, and physician and patient barriers to cancer clinical trial participation. *J Natl Cancer Inst* 2019 Mar 01;111(3):245-255 [FREE Full text] [doi: [10.1093/jnci/djy221](https://doi.org/10.1093/jnci/djy221)] [Medline: [30856272](https://pubmed.ncbi.nlm.nih.gov/30856272/)]
43. Unger JM, Cook E, Tai E, Bleyer A. The role of clinical trial participation in cancer research: barriers, evidence, and strategies. *Am Soc Clin Oncol Educ Book* 2016 May;35:185-198 [FREE Full text] [doi: [10.1200/EDBK_156686](https://doi.org/10.1200/EDBK_156686)] [Medline: [27249699](https://pubmed.ncbi.nlm.nih.gov/27249699/)]
44. Wedam S, Fashoyin-Aje L, Bloomquist E, Tang S, Sridhara R, Goldberg KB, et al. FDA approval summary: palbociclib for male patients with metastatic breast cancer. *Clin Cancer Res* 2020 Mar 15;26(6):1208-1212. [doi: [10.1158/1078-0432.CCR-19-2580](https://doi.org/10.1158/1078-0432.CCR-19-2580)] [Medline: [31649043](https://pubmed.ncbi.nlm.nih.gov/31649043/)]
45. Clinical trial eligibility design. Shiny Apps. URL: <https://zongzhi-liu.shinyapps.io/trialdesign/> [accessed 2024-06-18]
46. trialdesign. GitHub. URL: <https://github.com/zz2liu/trialdesign> [accessed 2024-06-18]

Abbreviations

- ALT:** alanine transaminase
- AST:** aspartate aminotransferase
- BCa:** breast cancer

CD: Crohn disease
ECOG: Eastern Cooperative Oncology Group
EGFR: epidermal growth factor receptor
EHR: electronic health record
LOINC: Logical Observation Identifiers, Names, and Codes
MM: multiple myeloma
NLP: natural language processing
NSCLC: non–small cell lung cancer
PCa: prostate cancer
SCA: synthetic control arm
UC: ulcerative colitis
ULN: upper limit of normal

Edited by K El Emam, B Malin; submitted 16.07.23; peer-reviewed by MO Khursheed, D Chrimes, P Han, L Guo; comments to author 18.10.23; revised version received 07.11.23; accepted 23.03.24; published 29.07.24.

Please cite as:

*Lee K, Liu Z, Mai Y, Jun T, Ma M, Wang T, Ai L, Calay E, Oh W, Stolovitzky G, Schadt E, Wang X
Optimizing Clinical Trial Eligibility Design Using Natural Language Processing Models and Real-World Data: Algorithm Development and Validation
JMIR AI 2024;3:e50800
URL: <https://ai.jmir.org/2024/1/e50800>
doi: [10.2196/50800](https://doi.org/10.2196/50800)
PMID: [39073872](https://pubmed.ncbi.nlm.nih.gov/39073872/)*

©Kyeryoung Lee, Zongzhi Liu, Yun Mai, Tomi Jun, Meng Ma, Tongyu Wang, Lei Ai, Ediz Calay, William Oh, Gustavo Stolovitzky, Eric Schadt, Xiaoyan Wang. Originally published in JMIR AI (<https://ai.jmir.org>), 29.07.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR AI, is properly cited. The complete bibliographic information, a link to the original publication on <https://www.ai.jmir.org/>, as well as this copyright and license information must be included.

Original Paper

A Comparison of Personalized and Generalized Approaches to Emotion Recognition Using Consumer Wearable Devices: Machine Learning Study

Joe Li¹; Peter Washington¹, PhD

Information and Computer Sciences, University of Hawai'i at Mānoa, Honolulu, HI, United States

Corresponding Author:

Peter Washington, PhD

Information and Computer Sciences

University of Hawai'i at Mānoa

1680 East-West Road, Room 312

Honolulu, HI, 96822

United States

Phone: 1 000000000

Email: pyw@hawaii.edu

Abstract

Background: There are a wide range of potential adverse health effects, ranging from headaches to cardiovascular disease, associated with long-term negative emotions and chronic stress. Because many indicators of stress are imperceptible to observers, the early detection of stress remains a pressing medical need, as it can enable early intervention. Physiological signals offer a noninvasive method for monitoring affective states and are recorded by a growing number of commercially available wearables.

Objective: We aim to study the differences between personalized and generalized machine learning models for 3-class emotion classification (neutral, stress, and amusement) using wearable biosignal data.

Methods: We developed a neural network for the 3-class emotion classification problem using data from the Wearable Stress and Affect Detection (WESAD) data set, a multimodal data set with physiological signals from 15 participants. We compared the results between a participant-exclusive generalized, a participant-inclusive generalized, and a personalized deep learning model.

Results: For the 3-class classification problem, our personalized model achieved an average accuracy of 95.06% and an F_1 -score of 91.71%; our participant-inclusive generalized model achieved an average accuracy of 66.95% and an F_1 -score of 42.50%; and our participant-exclusive generalized model achieved an average accuracy of 67.65% and an F_1 -score of 43.05%.

Conclusions: Our results emphasize the need for increased research in personalized emotion recognition models given that they outperform generalized models in certain contexts. We also demonstrate that personalized machine learning models for emotion classification are viable and can achieve high performance.

(JMIR AI 2024;3:e52171) doi:[10.2196/52171](https://doi.org/10.2196/52171)

KEYWORDS

affect detection; affective computing; deep learning; digital health; emotion recognition; machine learning; mental health; personalization; stress detection; wearable technology

Introduction

Stress and negative affect can have long-term consequences for physical and mental health, such as chronic illness, higher mortality rates, and major depression [1-3]. Therefore, the early detection and corresponding intervention of stress and negative emotions greatly reduces the risk of detrimental health conditions appearing later in life [4]. Since negative stress and

affect can be difficult for humans to observe [5-7], automated emotion recognition models can play an important role in health care. Affective computing can also facilitate digital therapy and advance the development of assistive technologies for autism [8-13].

Physiological signals, including electrocardiography (ECG), electrodermal activity (EDA), and photoplethysmography (PPG), have been shown to be robust indicators of emotions [14-16].

The noninvasive nature of physiological signal measurement makes it a practical and convenient method for emotion recognition. Wearable devices such as smartwatches have become increasingly popular, and products such as Fitbit have already integrated the sensing of heart rate, ECG, and EDA data into their smartwatches. The accessibility of wearable devices indicates that an emotion recognition model using biosignals can have practical applications in health care.

The vast majority of research in recognizing emotions from biosignals involves machine learning models that are generalizable, which means that the models were trained on one group of subjects and tested on a separate group of subjects [17-28]. Prior studies emphasize the need for personalized or subject-dependent models [18,29,30], and some investigations, albeit few, analyze personalized models [31,32]. Both generalized and personalized models have potential benefits; for example, generalized models can train on more data than personalized models, and personalized models do not need to address the problem of inter-subject data variance [33]. However, it is still unclear how personalized models compare against generalized models in many contexts.

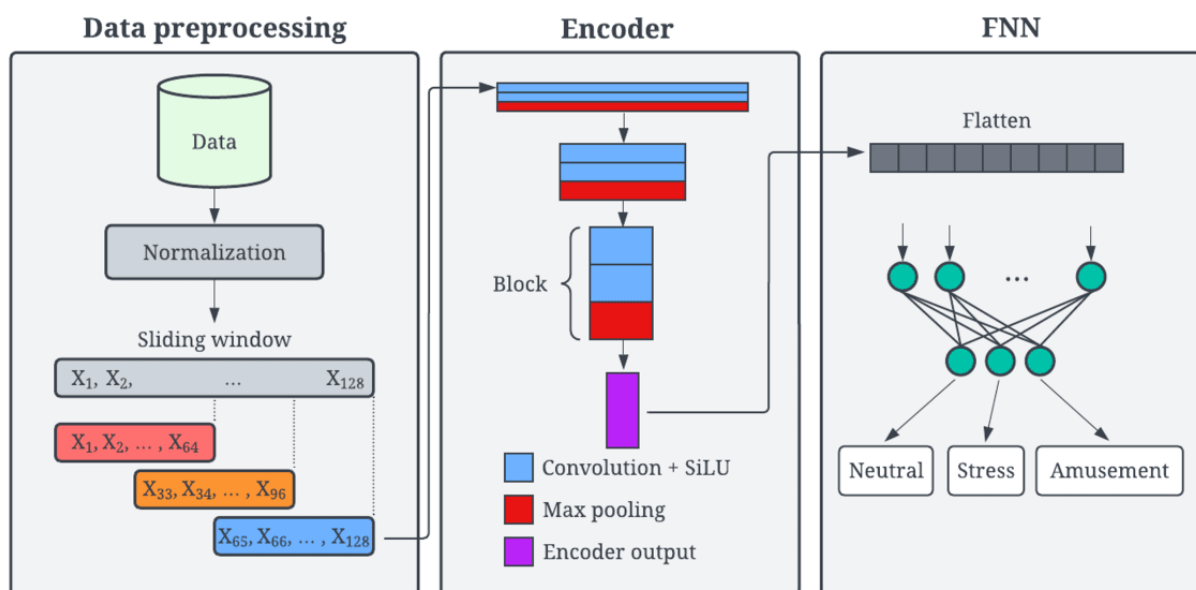
We present 1 personalized and 2 generalized machine learning approaches for the 3-class emotion classification problem (neutral, stress, and amusement) on the Wearable Stress and Affect Detection (WESAD) data set, a publicly available data set that includes both stress and emotion data [18]. The two generalized models are trained using participant-inclusive and participant-exclusive procedures. We compare the performance of these 3 models, finding that the personalized machine learning approach consistently outperforms the generalized approach on the WESAD data set.

Methods

Overview

To classify physiological data into the neutral, stress, and amusement classes, we developed a machine learning framework and evaluated the framework using data from the WESAD data set. Our machine learning framework consists of data preprocessing, a convolutional encoder for feature extraction, and a feedforward neural network for supervised prediction (Figure 1). Using this model architecture, we compared generalized and personalized approaches to the 3-class emotion classification task (neutral, stress, and amusement).

Figure 1. Overview of our model architecture for the 3-class emotion classification task. FNN: feedforward neural network; SiLU: sigmoid linear unit.



Data Set

We selected WESAD, a publicly available data set that combines both stress and emotion annotations. WESAD consists of multimodal physiological data in the form of continuous time-series data for 15 participants and corresponding annotations of 4 affective states: neutral, stress, amusement, and meditation. However, we only considered the neutral, stress, and amusement classes since the objective of WESAD is to provide data for the 3-class classification problem, and the benchmark model in WESAD ignores the meditation state as well. Our model incorporated data from 8 modalities recorded

in WESAD: ECG, EDA, electromyogram (EMG), respiration, temperature, and acceleration (x, y, and z axes). In the data set, measurements for each of the 8 modalities were sampled by a RespiBAN sensor at 700 Hz to enforce uniformity, and data were collected for approximately 36 minutes per participant.

Preprocessing and Partitioning

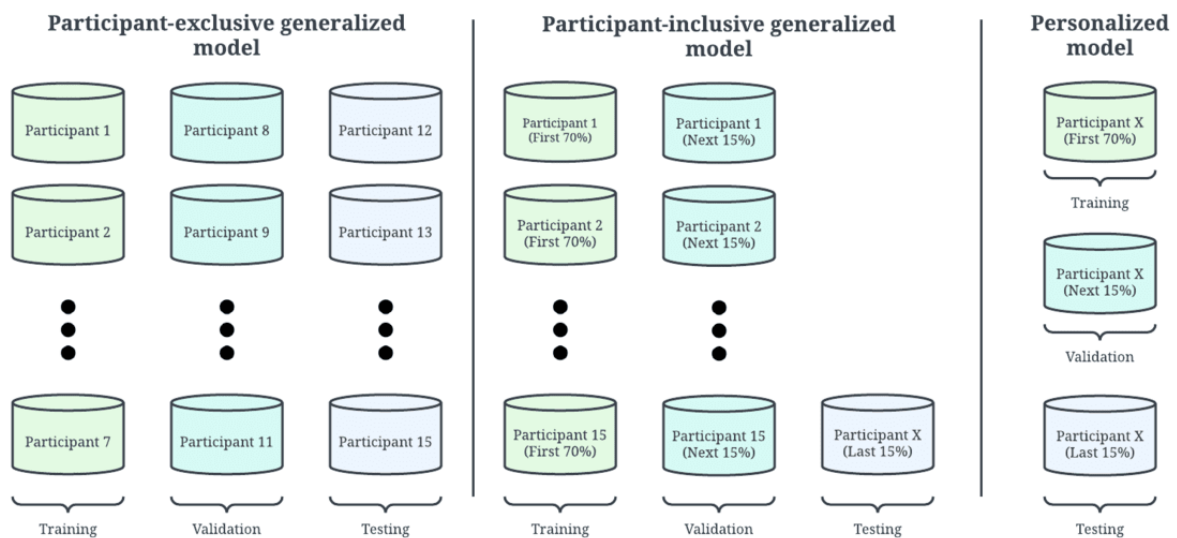
Each data modality was normalized with a mean of 0 and an SD of 1. We used a sliding window algorithm to partition each modality into intervals consisting of 64 data points, with a 50% overlap between consecutive intervals. We ensured that all 64 data points within an interval shared a common annotation,

which allowed us to assign a single affective state to each interval. The process of normalization, followed by a sliding window partition, is illustrated in Figure 1. These intervals were partitioned into training, validation, and testing sets.

For the personalized model, we partitioned the training, validation, and testing sets as follows: each participant in the data set had their own model that was trained, validated, and tested independently of other participants. For each affective state (neutral, stress, and amusement), we allocated the initial 70% of intervals with that affective state for training, the next

15% for validation, and the final 15% for testing. This guaranteed that the relative frequencies of each affective state were consistent across all 3 sets. Simply using the first 70% of all intervals for the training data would skew the distribution of affective states, given the nature of the WESAD data set. Furthermore, our partitioning of intervals according to sequential time order rather than random selection helped prevent overfitting by guaranteeing that 2 adjacent intervals with similar features would be in the same set. The partitioning of training, validation, and testing sets for the personalized model is shown in Figure 2.

Figure 2. A comparison of different generalized and personalized approaches to the 3-class emotion classification task. The participant-exclusive generalized model mimics generalized approaches used in other papers. The participant-exclusive generalized model shown in the figure differs from what we use in this paper.



Standard generalized models partition the training, validation, and testing sets by participant [18]. We denote these standard models as participant-exclusive generalized models, as shown in Figure 2. Through this partitioning method, it is impossible to compare the performances of generalized and personalized models since they are solving two separate tasks. Therefore, we present a modified participant-exclusive generalized model that solves the same task as the personalized model. The testing set for our participant-exclusive generalized model consisted of the last 15% of intervals for each affective state for 1 participant. The training set consisted of the first 70% of intervals for each affective state for all participants except the 1 participant in the testing set, and the validation set consisted of the next 15% of intervals for all participants except the 1 participant in the testing set. The training and testing sets for this approach contained data from mutually exclusive sets of participants; this is where the name of the model, participant-exclusive, is derived from. Since the testing sets for the participant-exclusive generalized and personalized models are equivalent, it is possible to compare generalized and personalized approaches. This participant-exclusive generalized model served as our first generalized model baseline.

A second generalized model baseline was created, called the participant-inclusive generalized model. Like the testing sets

for the participant-exclusive generalized and personalized models, the testing set for this model contained the last 15% of intervals for each affective state for a single participant. The training set consisted of the first 70% of intervals for each affective state for all participants, and the validation set consisted of the next 15%. The set of participants in the training and testing sets overlapped by 1 participant—the subject in the testing set—which is why this model is called the participant-inclusive generalized model. This is illustrated in Figure 2.

Model Architecture

The model architecture consisted of an encoder network followed by a feedforward head, which is shown in Figure 1. A total of 8 channels, representing the 8 modalities we used from WESAD, served as input into an encoder network, which was modeled after the encoder section of U-Net [34]. The encoder network had 3 blocks, with each block consisting of two 1D convolutional layers (kernel size of 3) followed by 1D max pooling (kernel size of 2). The output of each convolution operation was passed through a sigmoid linear unit (SiLU) activation function. Between each block, we doubled the number of channels and added a dropout layer (15%) to reduce overfitting. The output of the encoder was flattened and passed

through 2 fully connected layers with SiLU activation to produce a 3-class probability distribution. Table 1 shows the hyperparameters that determine the model structure. These were

consistent between the participant-exclusive generalized, participant-inclusive generalized, and personalized models.

Table 1. Hyperparameters relating to model structure.

Hyperparameter	Value
Encoder depth (number of blocks), n	3
Dropout rate, %	15
Number of fully connected layers, n	2
Convolutional kernel size, n	3
Max pooling kernel size, n	2
Activation function	SiLU ^a

^aSiLU: sigmoid linear unit.

Model Training

We trained the 2 generalized baseline models and the personalized model under the same hyperparameters to guarantee a fair comparison. Both models were trained with cross-entropy loss using AdamW optimization. All models were written using PyTorch [35]. Within 1000 epochs, models with the lowest validation loss were saved for testing. A Nvidia GeForce RTX 4090 GPU was used for training. A separate personalized model was trained for each of the 15 participants. The participant-exclusive generalized model was trained 15 times, and the participant-inclusive generalized model was trained once. For model comparison, all models were tested on each of the 15 participants.

Ethical Considerations

This study did not require institutional review board (IRB) review because we exclusively used a commonly analyzed publicly available data set. We did not work with any human subjects.

Results

For the 3-class emotion classification task (neutral, stress, and amusement), Tables 2 and 3 illustrate the accuracy and F_1 -score of the personalized and generalized models when tested on each of the 15 participants. We include F_1 -score, a balanced evaluation metric consisting of the harmonic mean of precision

and recall, to accommodate for the imbalanced class distribution in WESAD [18]. In order to guarantee a fair comparison between the models, they had the same random seeds for model initialization, and their architecture and hyperparameters were the same. The accuracy and F_1 -score for the personalized model exceeded those of the participant-inclusive generalized model for all participants except participant 1, and the personalized model outperformed the participant-exclusive generalized model in terms of accuracy and F_1 -score for all participants. The personalized models for participants 1 and 2 also indicate subpar performance compared to other participants, which we address in the Discussion section.

Table 4 shows the average and SD of the accuracies and F_1 -scores across all participants for the 3 models. We achieved an average accuracy of 95.06%, 66.95%, and 67.65% for the personalized, participant-inclusive generalized, and participant-exclusive generalized models, respectively. We also achieved an average F_1 -score of 91.72%, 42.50%, and 43.05% for the personalized, participant-inclusive generalized, and participant-exclusive generalized models, respectively. Observing the error margins in Table 4, the differences in accuracy and F_1 -score between the personalized model and both generalized models are statistically significant. As shown in Table 5, we evaluated the P values between each model type for accuracy and F_1 -score through pairwise 2-tailed t tests to determine statistical significance.

Table 2. A comparison of model accuracy between the personalized and generalized models.

Participant	Model accuracy, %		
	Personalized model	Participant-inclusive generalized model	Participant-exclusive generalized model
1	68.36	82.69	53.94
2	82.32	67.12	81.91
3	99.99	82.81	82.81
4	99.90	82.86	82.31
5	98.02	82.94	74.67
6	99.57	54.57	54.03
7	100.00	82.05	83.23
8	100.00	53.72	53.70
9	100.00	51.86	51.83
10	93.69	82.05	79.85
11	100.00	60.86	62.11
12	98.34	53.53	53.60
13	99.81	53.26	65.35
14	100.00	53.47	53.54
15	85.83	60.43	81.91

Table 3. A comparison of F_1 -score between the personalized and generalized models.

Participant	F_1 -score, %		
	Personalized model	Participant-inclusive generalized model	Participant-exclusive generalized model
1	58.14	61.91	23.36
2	58.88	44.55	58.53
3	99.98	62.05	62.05
4	99.87	61.95	61.50
5	96.87	61.99	54.74
6	99.35	24.94	23.59
7	100.00	61.16	62.09
8	100.00	23.38	23.29
9	100.00	22.85	22.89
10	94.29	61.04	59.23
11	100.00	38.27	40.15
12	97.40	26.79	26.90
13	99.75	24.47	44.63
14	100.00	23.93	24.09
15	71.28	38.26	58.71

Table 4. Average accuracy and F_1 -score of models across all participants.

Model type	Accuracy, mean (SD [%])	F_1 -score, mean (SD [%])
Personalized	95.06 (9.24)	91.72 (15.33)
Participant-inclusive generalized	66.95 (13.76)	42.50 (17.37)
Participant-exclusive generalized	67.65 (13.48)	43.05 (17.20)

Table 5. *P* values of accuracy and F_1 -score comparisons between model types.

Model comparison	<i>P</i> value for accuracy	<i>P</i> value for F_1 -score
Personalized versus participant-inclusive generalized	$P < .001$	$P < .001$
Personalized versus participant-exclusive generalized	$P < .001$	$P < .001$
Participant-inclusive generalized versus participant-exclusive generalized	.81	.88

Discussion

Principal Findings

We demonstrated that a personalized deep learning model outperforms a generalized model in both the accuracy and F_1 -score metrics for the 3-class emotion classification task. By establishing two generalized model baselines through the participant-inclusive and participant-exclusive models, we created an alternative approach to the standard generalization technique of separating the training and testing sets by participant, and as a result, we were able to compare personalized and generalized approaches. Our personalized model achieved an accuracy of 95.06% and an F_1 -score of 91.72%, while our participant-inclusive generalized model achieved an accuracy of 66.95% and an F_1 -score of 42.50% and our participant-exclusive generalized model achieved an accuracy of 67.65% and an F_1 -score of 43.05%.

Our work indicates that personalized models for emotion recognition should be further explored in the realm of health care. Machine learning methods for emotion classification are clearly viable and can achieve high accuracy, as shown by our personalized model. Furthermore, given that numerous wearable technologies collect physiological signals, data acquisition is both straightforward and noninvasive. Combined with the popularity of consumer wearable technology, it is feasible to scale emotion recognition systems. This can ultimately play a major role in the early detection of stress and negative emotions, thus serving as a preventative measure for serious health problems.

Comparison With Previous Work

Generalized Models

The vast majority of prior studies using WESAD developed generalized approaches to the emotion classification task. Schmidt et al [18], the pioneers of WESAD, created several feature extraction models and achieved accuracies up to 80% for the 3-class classification task. Huynh et al [22] developed a deep neural network, trained on WESAD wrist signals, to outperform past approaches by 8.22%. Albaladejo-González et al [36] achieved an F_1 -score of 88.89% using an unsupervised local outlier factor model and 99.03% using a supervised multilayer perceptron. Additionally, they analyzed the transfer learning capabilities of different models between the WESAD and SWELL-KW (SWELL knowledge work) [37] data sets. Ghosh et al [38] achieved 94.8% accuracy using WESAD chest data by encoding time-series data into Gramian Angular Field images and employing deep learning techniques. Bajpai et al [39] investigated the k-nearest neighbor algorithm to explore the tradeoff between performance and the total number of

nearest neighbors using WESAD. Through federated learning, Almadhor et al [40] achieved 86.82% accuracy on data in WESAD using a deep neural network. Behinaein et al [41] developed a novel transformer approach and achieved state-of-the-art performance using only one modality from WESAD.

Personalized Models

Sah and Ghasemzadeh [30] developed a generalized approach using a convolutional neural network using 1 modality from WESAD. For the 3-class classification problem, they achieved an average accuracy of 92.85%. They used the leave-one-subject-out (LOSO) analysis to highlight the need for personalization. Indikawati and Winiarti [31] directly developed a personalized approach for the 4-class classification problem in WESAD (neutral, stress, amusement, and meditation). Using different feature extraction machine learning models, they achieved accuracies ranging from 88%-99% for the 15 participants. Liu et al [32] developed a federated learning approach using data from WESAD with the goal of preserving user privacy. In doing so, they developed a personalized model as a baseline, which achieved an average accuracy of 90.2%. Nkurikiyeyezu et al [42] determined that personalized models (95.2% accuracy) outperform generalized models (42.5% accuracy) for the stress versus no-stress task. By running additional experiments to further understand how personalized models compare to generalized models for the 3-class emotion classification task and by developing participant-inclusive and participant-exclusive versions of the generalized models, our work concretely demonstrates how personalization outperforms generalization and thus supports the conclusions of Nkurikiyeyezu et al [42].

Limitations and Future Work

As shown in Tables 2 and 3, the performance of our personalized model deteriorates for participants 1 and 2. To analyze the lack of performance improvement of the personalized model for these 2 participants, we visualized the means and SDs of the different modalities for each emotion class. In Figures 3-5, we illustrate notable deviations in modality means and SDs for participants 1 and 2 compared to other participants. While the analysis of these modalities reveals important information about the nature of the WESAD data set, it still remains difficult to pinpoint the exact data set features that caused the performance decline in the personalized model for these 2 participants. This is another limitation: since we do not use a feature extraction model, we cannot assign a feature importance (eg, Gini importance) to individual features like Schmidt et al [18] do. We also analyzed the emotion class balances for each participant, which are included in Table 6, to see if anomalies existed in the class distributions for certain participants.

However, based on the ranges of the class distributions, class balance likely had minimal effect on the performance decline.

Figure 3. Deviations of mean and SD for participants 1 and 2 for neutral class modalities.

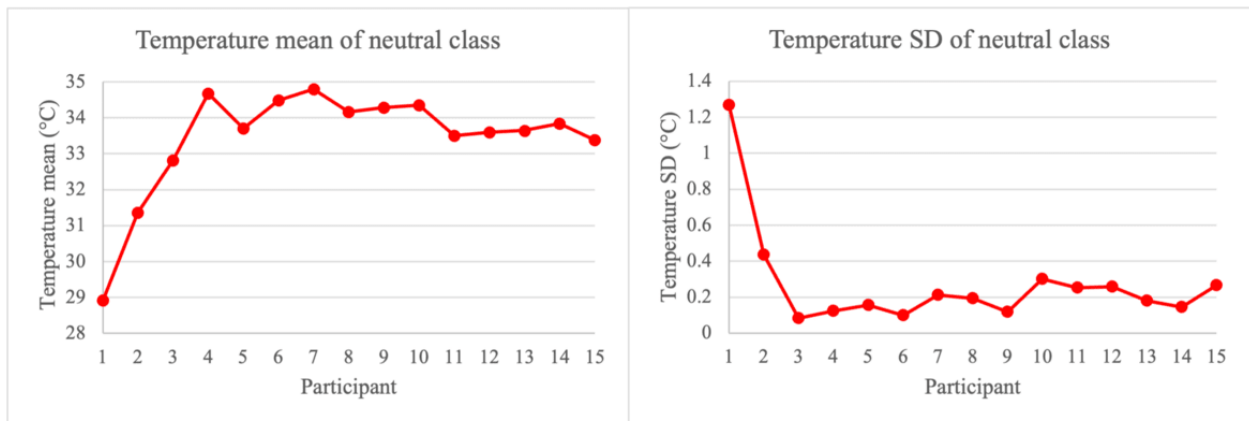


Figure 4. Deviations of mean and SD for subjects 1 and 2 for stress class modalities. EMG: electromyogram.

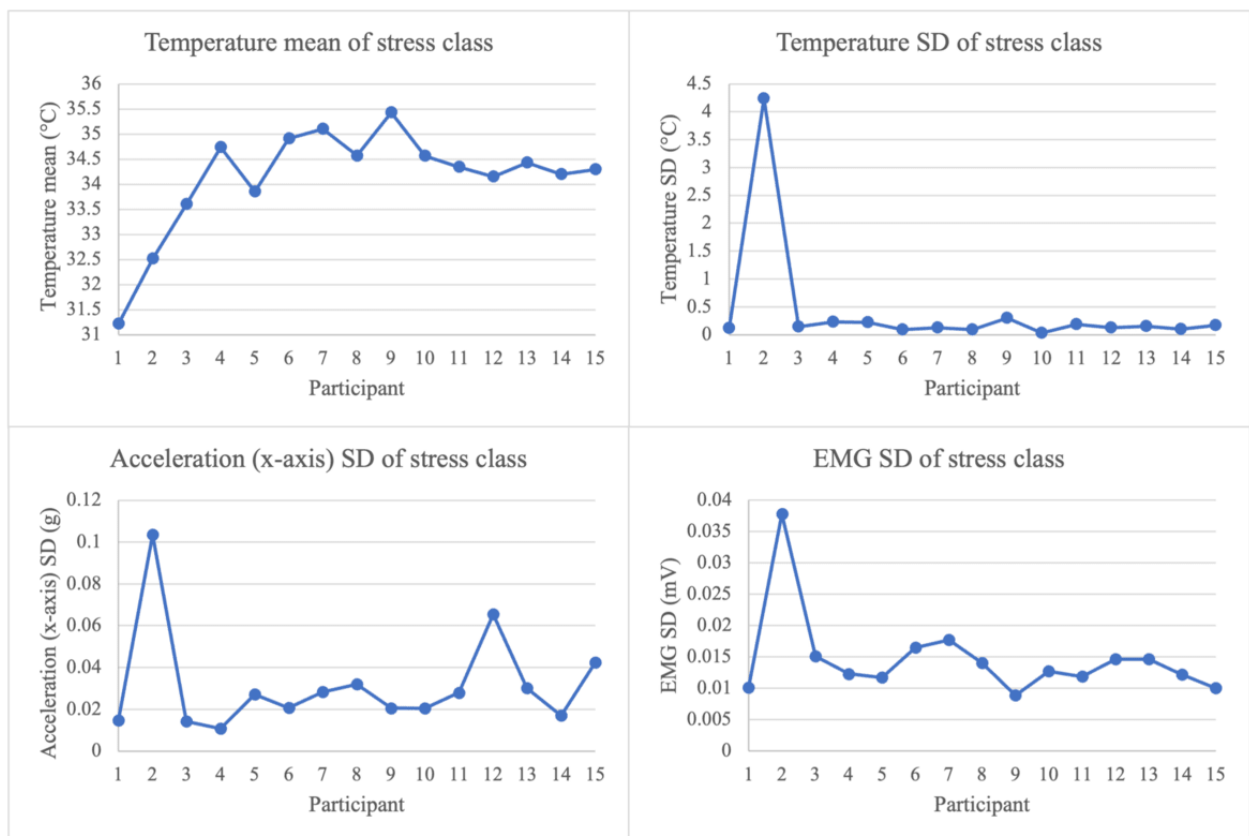
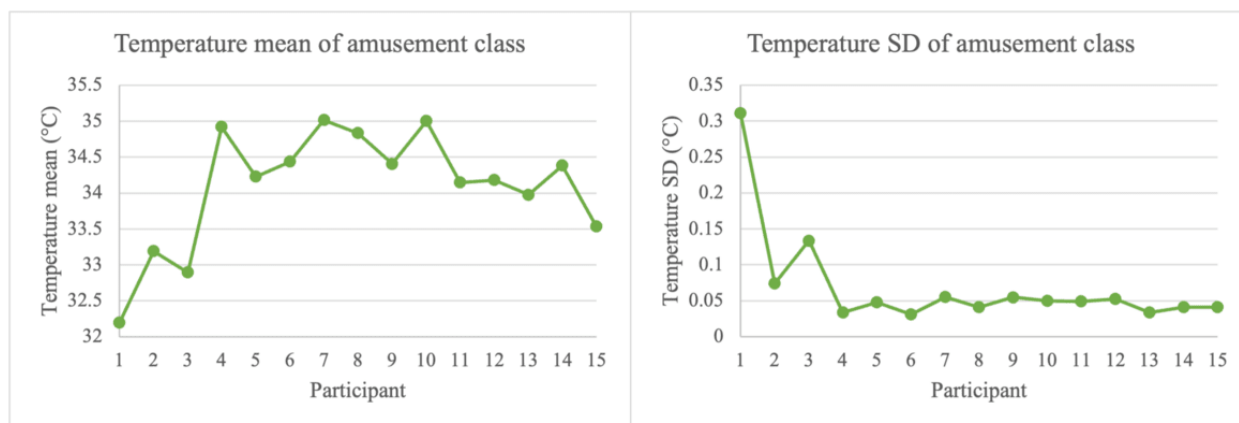


Figure 5. Deviations of mean and SD for subjects 1 and 2 for amusement class modalities.**Table 6.** Ranges of emotion class distributions per participant.

Emotion class	Range, %
Neutral	51.8-54.0
Stress	29.0-31.8
Amusement	16.3-17.4

Our participant-inclusive and participant-exclusive generalized models do not outperform previously published generalized models on the WESAD data set (eg, Schmidt et al [18] achieved up to 80% accuracy while we achieved 66.95% accuracy with our participant-inclusive model). This discrepancy can be attributed to a deliberate choice in our methodology: instead of maximizing our generalized models' performance with hyperparameter tuning, we simply opted for a consistent set of hyperparameters across the personalized and generalized models because our primary objective was to evaluate their relative performance. While hyperparameter tuning might yield higher results in practice, differing hyperparameters between our models would introduce additional variables that make it difficult to determine the role that personalization and generalization play in model performance.

Given the variations between participants, one approach to improving generalized model performance is adding embedding representations for each participant or participant-specific demographic data as additional features as a method of distinguishing individual participants in generalized models. However, to prevent overfitting to participant-specific features like demographic data, data sets with significantly more participants would need to be created, given the small sample size of the WESAD data set.

One limitation that personalized models may encounter during training is the cold start problem, given that personalized models receive less data than generalized models. Moreover, despite the accuracy improvement in personalized models, developing a model for each participant may be costly and unscalable: data must be labeled specifically per participant, and enough data must be provided to the model to overcome the cold start problem (notably, however, even though the cold start problem should theoretically put our personalized model at a disadvantage, the WESAD data set provided enough data for

our personalized model to outperform our generalized model). Both of these limitations can be addressed by a self-supervised learning approach to emotion recognition.

A self-supervised learning approach follows a framework used by natural language processing models such as the Bidirectional Encoder Representations from Transformers (BERT) model [43]. A model first pretrains on a large set of unlabeled data across numerous participants. Then, the pretrained model is fine-tuned to a small amount of labeled, participant-specific data. The pretraining phase eliminates the burden of manual labeling because all data are unlabeled, as well as the cold start problem because large amounts of data can be provided. The fine-tuning phase requires only a small amount of user-specific labeled data to perform accurately, and studies have already begun exploring the tradeoffs between the number of labels and model accuracy in WESAD using self-supervised or semisupervised approaches [44,45].

Finally, to expand beyond the WESAD data set, it is valuable to reproduce results on additional physiological signal data sets for emotion analysis, such as the Database for Emotion Analysis using Physiological Signals (DEAP) [46] and Cognitive Load, Affect, and Stress (CLAS) [47]. Data from WESAD were collected under controlled laboratory environments, which may not generalize to the real world. Therefore, analyzing emotions in a real-world context through data sets such as K-EmoCon [48], which contain physiological data collected in naturalistic conversations, may be useful. Emotions in the K-EmoCon data set were categorized into 18 different classes, so exploring this data set could also help us better assess the benefits of personalization for a broader range of emotions. A major goal of this approach is to provide support for personalized digital interventions for neuropsychiatry, which could benefit a variety of applications, such as video-based digital therapeutics for

children with autism to predict the child's affective state as part of the therapeutic process [49-52].

Acknowledgments

The project described was supported by grant U54GM138062 from the National Institute of General Medical Sciences (NIGMS), a component of the National Institutes of Health (NIH), and its contents are solely the responsibility of the author and do not necessarily represent the official view of NIGMS or NIH. The project was also supported by a grant from the Medical Research Award fund of the Hawai'i Community Foundation (grant MedRes_2023_00002689).

Conflicts of Interest

None declared.

References

1. Kendler KS, Karkowski LM, Prescott CA. Causal relationship between stressful life events and the onset of major depression. *Am J Psychiatry* 1999;156(6):837-841 [FREE Full text] [doi: [10.1176/ajp.156.6.837](https://doi.org/10.1176/ajp.156.6.837)] [Medline: [10360120](https://pubmed.ncbi.nlm.nih.gov/10360120/)]
2. Chiang JJ, Turiano NA, Mroczek DK, Miller GE. Affective reactivity to daily stress and 20-year mortality risk in adults with chronic illness: findings from the national study of daily experiences. *Health Psychol* 2018;37(2):170-178 [FREE Full text] [doi: [10.1037/hea0000567](https://doi.org/10.1037/hea0000567)] [Medline: [29154603](https://pubmed.ncbi.nlm.nih.gov/29154603/)]
3. Leger KA, Charles ST, Almeida DM. Let it go: lingering negative affect in response to daily stressors is associated with physical health years later. *Psychol Sci* 2018;29(8):1283-1290 [FREE Full text] [doi: [10.1177/0956797618763097](https://doi.org/10.1177/0956797618763097)] [Medline: [29553880](https://pubmed.ncbi.nlm.nih.gov/29553880/)]
4. Jorm AF. Mental health literacy: empowering the community to take action for better mental health. *Am Psychol* 2012;67(3):231-243. [doi: [10.1037/a0025957](https://doi.org/10.1037/a0025957)] [Medline: [22040221](https://pubmed.ncbi.nlm.nih.gov/22040221/)]
5. Mauss IB, Cook CL, Cheng JYJ, Gross JJ. Individual differences in cognitive reappraisal: experiential and physiological responses to an anger provocation. *Int J Psychophysiol* 2007;66(2):116-124. [doi: [10.1016/j.ijpsycho.2007.03.017](https://doi.org/10.1016/j.ijpsycho.2007.03.017)] [Medline: [17543404](https://pubmed.ncbi.nlm.nih.gov/17543404/)]
6. Jordan AH, Monin B, Dweck CS, Lovett BJ, John OP, Gross JJ. Misery has more company than people think: underestimating the prevalence of others' negative emotions. *Pers Soc Psychol Bull* 2011;37(1):120-135 [FREE Full text] [doi: [10.1177/0146167210390822](https://doi.org/10.1177/0146167210390822)] [Medline: [21177878](https://pubmed.ncbi.nlm.nih.gov/21177878/)]
7. Lane RD, Smith R. Levels of emotional awareness: theory and measurement of a socio-emotional skill. *J Intell* 2021;9(3):42 [FREE Full text] [doi: [10.3390/jintelligence9030042](https://doi.org/10.3390/jintelligence9030042)] [Medline: [34449662](https://pubmed.ncbi.nlm.nih.gov/34449662/)]
8. el Kaliouby R, Picard R, Baron-Cohen S. Affective computing and autism. *Ann N Y Acad Sci* 2006;1093:228-248. [doi: [10.1196/annals.1382.016](https://doi.org/10.1196/annals.1382.016)] [Medline: [17312261](https://pubmed.ncbi.nlm.nih.gov/17312261/)]
9. D'Alfonso S, Lederman R, Bucci S, Berry K. The digital therapeutic alliance and human-computer interaction. *JMIR Ment Health* 2020;7(12):e21895 [FREE Full text] [doi: [10.2196/21895](https://doi.org/10.2196/21895)] [Medline: [33372897](https://pubmed.ncbi.nlm.nih.gov/33372897/)]
10. Washington P, Wall DP. A review of and roadmap for data science and machine learning for the neuropsychiatric phenotype of autism. *Annu Rev Biomed Data Sci* 2023;6:211-228 [FREE Full text] [doi: [10.1146/annurev-biodatasci-020722-125454](https://doi.org/10.1146/annurev-biodatasci-020722-125454)] [Medline: [37137169](https://pubmed.ncbi.nlm.nih.gov/37137169/)]
11. Washington P, Park N, Srivastava P, Voss C, Kline A, Varma M, et al. Data-driven diagnostics and the potential of mobile artificial intelligence for digital therapeutic phenotyping in computational psychiatry. *Biol Psychiatry Cogn Neurosci Neuroimaging* 2020;5(8):759-769 [FREE Full text] [doi: [10.1016/j.bpsc.2019.11.015](https://doi.org/10.1016/j.bpsc.2019.11.015)] [Medline: [32085921](https://pubmed.ncbi.nlm.nih.gov/32085921/)]
12. Voss C, Schwartz J, Daniels J, Kline A, Haber N, Washington P, et al. Effect of wearable digital intervention for improving socialization in children with autism spectrum disorder: a randomized clinical trial. *JAMA Pediatr* 2019;173(5):446-454 [FREE Full text] [doi: [10.1001/jamapediatrics.2019.0285](https://doi.org/10.1001/jamapediatrics.2019.0285)] [Medline: [30907929](https://pubmed.ncbi.nlm.nih.gov/30907929/)]
13. Washington P, Voss C, Kline A, Haber N, Daniels J, Fazel A, et al. SuperpowerGlass: a wearable aid for the at-home therapy of children with autism. *Proc ACM Interact Mob Wearable Ubiquitous Technol* 2017;1(3):1-22. [doi: [10.1145/3130977](https://doi.org/10.1145/3130977)]
14. Rainville P, Bechara A, Naqvi N, Damasio AR. Basic emotions are associated with distinct patterns of cardiorespiratory activity. *Int J Psychophysiol* 2006;61(1):5-18. [doi: [10.1016/j.ijpsycho.2005.10.024](https://doi.org/10.1016/j.ijpsycho.2005.10.024)] [Medline: [16439033](https://pubmed.ncbi.nlm.nih.gov/16439033/)]
15. Nummenmaa L, Glerean E, Hari R, Hietanen JK. Bodily maps of emotions. *Proc Natl Acad Sci U S A* 2014;111(2):646-651 [FREE Full text] [doi: [10.1073/pnas.1321664111](https://doi.org/10.1073/pnas.1321664111)] [Medline: [24379370](https://pubmed.ncbi.nlm.nih.gov/24379370/)]
16. Jang EH, Park BJ, Park MS, Kim SH, Sohn JH. Analysis of physiological signals for recognition of boredom, pain, and surprise emotions. *J Physiol Anthropol* 2015;34(1):25 [FREE Full text] [doi: [10.1186/s40101-015-0063-5](https://doi.org/10.1186/s40101-015-0063-5)] [Medline: [26084816](https://pubmed.ncbi.nlm.nih.gov/26084816/)]
17. Domínguez-Jiménez JA, Campo-Landines KC, Martínez-Santos JC, Delahoz EJ, Contreras-Ortiz SH. A machine learning model for emotion recognition from physiological signals. *Biomed Signal Process Control* 2020;55:101646. [doi: [10.1016/j.bspc.2019.101646](https://doi.org/10.1016/j.bspc.2019.101646)]

18. Schmidt P, Reiss A, Duerichen R, Marberger C, Van Laerhoven K. Introducing WESAD, a multimodal dataset for wearable stress and affect detection. 2018 Presented at: ICMI '18: Proceedings of the 20th ACM International Conference on Multimodal Interaction; October 16-20, 2018; Boulder, CO p. 400-408. [doi: [10.1145/3242969.3242985](https://doi.org/10.1145/3242969.3242985)]
19. He C, Yao YJ, Ye XS. An emotion recognition system based on physiological signals obtained by wearable sensors. In: Yang C, Virk GS, Yang H, editors. *Wearable Sensors and Robots: Proceedings of International Conference on Wearable Sensors and Robots 2015*. Singapore: Springer; 2017:15-25.
20. Ramzan M, Dawn S. Fused CNN-LSTM deep learning emotion recognition model using electroencephalography signals. *Int J Neurosci* 2023;133(6):587-597. [doi: [10.1080/00207454.2021.1941947](https://doi.org/10.1080/00207454.2021.1941947)] [Medline: [34121598](https://pubmed.ncbi.nlm.nih.gov/34121598/)]
21. Vijayakumar S, Flynn R, Murray N. A comparative study of machine learning techniques for emotion recognition from peripheral physiological signals. 2020 Presented at: 2020 31st Irish Signals and Systems Conference (ISSC); June 11-12, 2020; Letterkenny, Ireland. [doi: [10.1109/issc49989.2020.9180193](https://doi.org/10.1109/issc49989.2020.9180193)]
22. Huynh L, Nguyen T, Nguyen T, Pirttikangas S, Siirtola P. StressNAS: affect state and stress detection using neural architecture search. 2021 Presented at: UbiComp/ISWC '21 Adjunct: Adjunct Proceedings of the 2021 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2021 ACM International Symposium on Wearable Computers; September 21-26, 2021; Virtual p. 121-125. [doi: [10.1145/3460418.3479320](https://doi.org/10.1145/3460418.3479320)]
23. Hsieh CP, Chen YT, Beh WK, Wu AYA. Feature selection framework for XGBoost based on electrodermal activity in stress detection. 2019 Presented at: 2019 IEEE International Workshop on Signal Processing Systems (SiPS); October 20-23, 2019; Nanjing, China. [doi: [10.1109/sips47522.2019.9020321](https://doi.org/10.1109/sips47522.2019.9020321)]
24. Garg P, Santhosh J, Dengel A, Ishimaru S. Stress detection by machine learning and wearable sensors. 2021 Presented at: IUI '21 Companion: 26th International Conference on Intelligent User Interfaces - Companion; April 14-17, 2021; College Station, TX p. 43-45. [doi: [10.1145/3397482.3450732](https://doi.org/10.1145/3397482.3450732)]
25. Lai K, Yanushkevich SN, Shmerko VP. Intelligent stress monitoring assistant for first responders. *IEEE Access* 2021;9:25314-25329 [FREE Full text] [doi: [10.1109/access.2021.3057578](https://doi.org/10.1109/access.2021.3057578)]
26. Siirtola P. Continuous stress detection using the sensors of commercial smartwatch. 2019 Presented at: UbiComp/ISWC '19 Adjunct: Adjunct Proceedings of the 2019 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2019 ACM International Symposium on Wearable Computers; September 9-13, 2019; London, United Kingdom p. 1198-1201. [doi: [10.1145/3341162.3344831](https://doi.org/10.1145/3341162.3344831)]
27. Bobade P, Vani M. Stress detection with machine learning and deep learning using multimodal physiological data. 2020 Presented at: 2020 Second International Conference on Inventive Research in Computing Applications (ICIRCA); July 15-17, 2020; Coimbatore, India. [doi: [10.1109/icirca48905.2020.9183244](https://doi.org/10.1109/icirca48905.2020.9183244)]
28. Kumar A, Sharma K, Sharma A. Hierarchical deep neural network for mental stress state detection using IoT based biomarkers. *Pattern Recognit Lett* 2021;145:81-87. [doi: [10.1016/j.patrec.2021.01.030](https://doi.org/10.1016/j.patrec.2021.01.030)]
29. Schmidt P, Reiss A, Dürichen R, Van Laerhoven K. Wearable-based affect recognition: a review. *Sensors (Basel)* 2019;19(19):4079 [FREE Full text] [doi: [10.3390/s19194079](https://doi.org/10.3390/s19194079)] [Medline: [31547220](https://pubmed.ncbi.nlm.nih.gov/31547220/)]
30. Sah RK, Ghasemzadeh H. Stress classification and personalization: getting the most out of the least. *ArXiv Preprint* posted online on July 12 2021. [doi: [10.48550/arXiv.2107.05666](https://doi.org/10.48550/arXiv.2107.05666)]
31. Indikawati FI, Winiarti S. Stress detection from multimodal wearable sensor data. *IOP Conf Ser Mater Sci Eng* 2020;771(1):012028 [FREE Full text] [doi: [10.1088/1757-899X/771/1/012028](https://doi.org/10.1088/1757-899X/771/1/012028)]
32. Liu JC, Goetz J, Sen S, Tewari A. Learning from others without sacrificing privacy: simulation comparing centralized and federated machine learning on mobile health data. *JMIR Mhealth Uhealth* 2021;9(3):e23728 [FREE Full text] [doi: [10.2196/23728](https://doi.org/10.2196/23728)] [Medline: [33783362](https://pubmed.ncbi.nlm.nih.gov/33783362/)]
33. Ahmad Z, Khan N. A survey on physiological signal-based emotion recognition. *Bioengineering (Basel)* 2022;9(11):688 [FREE Full text] [doi: [10.3390/bioengineering9110688](https://doi.org/10.3390/bioengineering9110688)] [Medline: [36421089](https://pubmed.ncbi.nlm.nih.gov/36421089/)]
34. Ronneberger O, Fischer P, Brox T. U-net: convolutional networks for biomedical image segmentation. In: Navab N, Hornegger J, Wells WM, Frangi AF, editors. *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III*. Cham, Switzerland: Springer; 2015:234-241.
35. Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, et al. PyTorch: an imperative style, high-performance deep learning library. 2019 Presented at: 33rd Conference on Neural Information Processing Systems (NeurIPS 2019); December 8–14, 2019; Vancouver, BC.
36. Albaladejo-González M, Ruipérez-Valiente JA, Gómez Mármol F. Evaluating different configurations of machine learning models and their transfer learning capabilities for stress detection using heart rate. *J Ambient Intell Humaniz Comput* 2023;14(8):11011-11021 [FREE Full text] [doi: [10.1007/s12652-022-04365-z](https://doi.org/10.1007/s12652-022-04365-z)]
37. Koldijk S, Sappelli M, Verberne S, Neerincx MA, Kraaij W. The SWELL knowledge work dataset for stress and user modeling research. 2014 Presented at: ICMI '14: Proceedings of the 16th International Conference on Multimodal Interaction; November 12-16, 2014; Istanbul, Turkey p. 291-298. [doi: [10.1145/2663204.2663257](https://doi.org/10.1145/2663204.2663257)]
38. Ghosh S, Kim S, Ijaz MF, Singh PK, Mahmud M. Classification of mental stress from wearable physiological sensors using image-encoding-based deep neural network. *Biosensors (Basel)* 2022;12(12):1153 [FREE Full text] [doi: [10.3390/bios12121153](https://doi.org/10.3390/bios12121153)] [Medline: [36551120](https://pubmed.ncbi.nlm.nih.gov/36551120/)]

39. Bajpai D, He L. Evaluating KNN performance on WESAD dataset. 2020 Presented at: 2020 12th International Conference on Computational Intelligence and Communication Networks (CICN); September 25-26, 2020; Bhimtal, India. [doi: [10.1109/cicn49253.2020.9242568](https://doi.org/10.1109/cicn49253.2020.9242568)]
40. Almadhor A, Sampedro GA, Abisado M, Abbas S, Kim YJ, Khan MA, et al. Wrist-based electrodermal activity monitoring for stress detection using federated learning. *Sensors (Basel)* 2023;23(8):3984 [FREE Full text] [doi: [10.3390/s23083984](https://doi.org/10.3390/s23083984)] [Medline: [37112323](https://pubmed.ncbi.nlm.nih.gov/37112323/)]
41. Behinaein B, Bhatti A, Rodenburg D, Hungler P, Etemad A. A transformer architecture for stress detection from ECG. 2021 Presented at: ISWC '21: Proceedings of the 2021 ACM International Symposium on Wearable Computers; September 21-26, 2021; Virtual p. 132-134. [doi: [10.1145/3460421.3480427](https://doi.org/10.1145/3460421.3480427)]
42. Nkurikiyeyezu K, Yokokubo A, Lopez G. The effect of person-specific biometrics in improving generic stress predictive models. *ArXiv Preprint* posted online on December 31 2019. [doi: [10.48550/arXiv.1910.01770](https://doi.org/10.48550/arXiv.1910.01770)]
43. Devlin J, Chang MW, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. *ArXiv Preprint* posted online on May 24 2019. [doi: [10.48550/arXiv.1810.04805](https://doi.org/10.48550/arXiv.1810.04805)]
44. Khan N, Sarkar N. Semi-supervised generative adversarial network for stress detection using partially labeled physiological data. *ArXiv Preprint* posted online on October 27 2022. [doi: [10.48550/arXiv.2206.14976](https://doi.org/10.48550/arXiv.2206.14976)]
45. Islam T, Washington P. Personalized prediction of recurrent stress events using self-supervised learning on multimodal time-series data. *ArXiv Preprint* posted online on July 07 2023. [doi: [10.48550/arXiv.2307.03337](https://doi.org/10.48550/arXiv.2307.03337)]
46. Koelstra S, Muhl C, Soleymani M, Lee JS, Yazdani A, Ebrahimi T, et al. DEAP: a database for emotion analysis using physiological signals. *IEEE Trans Affect Comput* 2012;3(1):18-31. [doi: [10.1109/t-affc.2011.15](https://doi.org/10.1109/t-affc.2011.15)]
47. Markova V, Ganchev T, Kalinkov K. CLAS: a database for cognitive load, affect and stress recognition. 2019 Presented at: 2019 International Conference on Biomedical Innovations and Applications (BIA); November 8-9, 2019; Varna, Bulgaria. [doi: [10.1109/bia48344.2019.8967457](https://doi.org/10.1109/bia48344.2019.8967457)]
48. Park CY, Cha N, Kang S, Kim A, Khandoker AH, Hadjileontiadis L, et al. K-EmoCon, a multimodal sensor dataset for continuous emotion recognition in naturalistic conversations. *Sci Data* 2020;7(1):293 [FREE Full text] [doi: [10.1038/s41597-020-00630-y](https://doi.org/10.1038/s41597-020-00630-y)] [Medline: [32901038](https://pubmed.ncbi.nlm.nih.gov/32901038/)]
49. Daniels J, Schwartz JN, Voss C, Haber N, Fazel A, Kline A, et al. Exploratory study examining the at-home feasibility of a wearable tool for social-affective learning in children with autism. *NPJ Digit Med* 2018;1(1):32 [FREE Full text] [doi: [10.1038/s41746-018-0035-3](https://doi.org/10.1038/s41746-018-0035-3)] [Medline: [31304314](https://pubmed.ncbi.nlm.nih.gov/31304314/)]
50. Daniels J, Haber N, Voss C, Schwartz J, Tamura S, Fazel A, et al. Feasibility testing of a wearable behavioral aid for social learning in children with autism. *Appl Clin Inform* 2018;9(1):129-140 [FREE Full text] [doi: [10.1055/s-0038-1626727](https://doi.org/10.1055/s-0038-1626727)] [Medline: [29466819](https://pubmed.ncbi.nlm.nih.gov/29466819/)]
51. Kalantarian H, Jedoui K, Washington P, Tariq Q, Dunlap K, Schwartz J, et al. Labeling images with facial emotion and the potential for pediatric healthcare. *Artif Intell Med* 2019;98:77-86 [FREE Full text] [doi: [10.1016/j.artmed.2019.06.004](https://doi.org/10.1016/j.artmed.2019.06.004)] [Medline: [31521254](https://pubmed.ncbi.nlm.nih.gov/31521254/)]
52. Kalantarian H, Jedoui K, Washington P, Wall DP. A mobile game for automatic emotion-labeling of images. *IEEE Trans Games* 2020;12(2):213-218 [FREE Full text] [doi: [10.1109/tg.2018.2877325](https://doi.org/10.1109/tg.2018.2877325)] [Medline: [32551410](https://pubmed.ncbi.nlm.nih.gov/32551410/)]

Abbreviations

- BERT:** Bidirectional Encoder Representations from Transformers
- CLAS:** Cognitive Load, Affect, and Stress
- DEAP:** Database for Emotion Analysis using Physiological Signals
- ECG:** electrocardiography
- EDA:** electrodermal activity
- EMG:** electromyogram
- LOSO:** leave-one-subject-out
- PPG:** photoplethysmography
- SiLU:** sigmoid linear unit
- SWELL:** Smart Reasoning for Well-being at Home and at Work
- SWELL-KW:** SWELL knowledge work
- WESAD:** Wearable Stress and Affect Dataset

Edited by K El Emam, B Malin; submitted 25.08.23; peer-reviewed by S Pandey, M Zhou, G Vos; comments to author 19.09.23; revised version received 19.02.24; accepted 23.03.24; published 10.05.24.

Please cite as:

Li J, Washington P

A Comparison of Personalized and Generalized Approaches to Emotion Recognition Using Consumer Wearable Devices: Machine Learning Study

JMIR AI 2024;3:e52171

URL: <https://ai.jmir.org/2024/1/e52171>

doi: [10.2196/52171](https://doi.org/10.2196/52171)

PMID: [38875573](https://pubmed.ncbi.nlm.nih.gov/38875573/)

©Joe Li, Peter Washington. Originally published in JMIR AI (<https://ai.jmir.org>), 10.05.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR AI, is properly cited. The complete bibliographic information, a link to the original publication on <https://www.ai.jmir.org/>, as well as this copyright and license information must be included.

Original Paper

Sample Size Considerations for Fine-Tuning Large Language Models for Named Entity Recognition Tasks: Methodological Study

Zoltan P Majdik¹, PhD; S Scott Graham², PhD; Jade C Shiva Edward², MA; Sabrina N Rodriguez³, BS; Martha S Karnes⁴, PhD; Jared T Jensen², MA; Joshua B Barbour⁵, PhD; Justin F Rousseau^{6,7}, MD, MMSc

¹Department of Communication, North Dakota State University, Fargo, ND, United States

²Department of Rhetoric & Writing, The University of Texas at Austin, Austin, TX, United States

³Department of Neurology, The Dell Medical School, The University of Texas at Austin, Austin, TX, United States

⁴Department of Rhetoric & Writing, University of Arkansas Little Rock, Little Rock, AR, United States

⁵Department of Communication, The University of Illinois at Urbana-Champaign, Urbana, IL, United States

⁶Statistical Planning and Analysis Section, Department of Neurology, The University of Texas Southwestern Medical Center, Dallas, TX, United States

⁷Peter O'Donnell Jr. Brain Institute, The University of Texas Southwestern Medical Center, Dallas, TX, United States

Corresponding Author:

S Scott Graham, PhD

Department of Rhetoric & Writing

The University of Texas at Austin

Parlin Hall 29

Mail Code: B5500

Austin, TX, 78712

United States

Phone: 1 512 475 9507

Email: ssg@utexas.edu

Abstract

Background: Large language models (LLMs) have the potential to support promising new applications in health informatics. However, practical data on sample size considerations for fine-tuning LLMs to perform specific tasks in biomedical and health policy contexts are lacking.

Objective: This study aims to evaluate sample size and sample selection techniques for fine-tuning LLMs to support improved named entity recognition (NER) for a custom data set of conflicts of interest disclosure statements.

Methods: A random sample of 200 disclosure statements was prepared for annotation. All “PERSON” and “ORG” entities were identified by each of the 2 raters, and once appropriate agreement was established, the annotators independently annotated an additional 290 disclosure statements. From the 490 annotated documents, 2500 stratified random samples in different size ranges were drawn. The 2500 training set subsamples were used to fine-tune a selection of language models across 2 model architectures (Bidirectional Encoder Representations from Transformers [BERT] and Generative Pre-trained Transformer [GPT]) for improved NER, and multiple regression was used to assess the relationship between sample size (sentences), entity density (entities per sentence [EPS]), and trained model performance (F_1 -score). Additionally, single-predictor threshold regression models were used to evaluate the possibility of diminishing marginal returns from increased sample size or entity density.

Results: Fine-tuned models ranged in topline NER performance from F_1 -score=0.79 to F_1 -score=0.96 across architectures.

Two-predictor multiple linear regression models were statistically significant with multiple R^2 ranging from 0.6057 to 0.7896 (all $P < .001$). EPS and the number of sentences were significant predictors of F_1 -scores in all cases ($P < .001$), except for the GPT-2_large model, where EPS was not a significant predictor ($P = .184$). Model thresholds indicate points of diminishing marginal return from increased training data set sample size measured by the number of sentences, with point estimates ranging from 439 sentences for RoBERTa_large to 527 sentences for GPT-2_large. Likewise, the threshold regression models indicate a diminishing marginal return for EPS with point estimates between 1.36 and 1.38.

Conclusions: Relatively modest sample sizes can be used to fine-tune LLMs for NER tasks applied to biomedical text, and training data entity density should representatively approximate entity density in production data. Training data quality and a

model architecture's intended use (text generation vs text processing or classification) may be as, or more, important as training data volume and model parameter size.

(*JMIR AI 2024;3:e52095*) doi:[10.2196/52095](https://doi.org/10.2196/52095)

KEYWORDS

named-entity recognition; large language models; fine-tuning; transfer learning; expert annotation; annotation; sample size; sample; language model; machine learning; natural language processing; disclosure; disclosures; statement; statements; conflict of interest

Introduction

Background

Named entity recognition (NER) has many applications in biomedical and clinical natural language processing (NLP). As its core function, NER identifies and categorizes specific terms or phrases representing people, places, organizations, and other entities. It has been used to identify or extract named entities in free-text clinical notes and reports in the secondary analysis of electronic health records [1,2]. NER has also been used alone or as part of an NLP pipeline to detect protected health information in order to deidentify clinical text for secondary analysis [3,4]. Additionally, NER has been used to identify and classify medications [5,6], specific disease and clinical condition entities [7], and laboratory tests [8] into existing taxonomies for purposes of secondary research, cohort generation, or clinical decision support [9-12]. While NER solutions have a long history of applications in NLP and clinical NLP domains, their effectiveness has recently been enhanced through the addition of large language models (LLMs) in relevant data parsing pipelines. LLMs have become an integral part of research pipelines in fields as diverse as digital humanities [13], computational social science [14], bioinformatics, applied ethics, and finance.

LLMs, such as GPT-3, have demonstrated remarkable performance across a variety of tasks. For instance, the GPT-3.5-powered LLM application ChatGPT performed close to or at the passing threshold of 60% accuracy on the United States Medical Licensing Exam (USMLE) without the specialized input of human trainers [15]. Widely available models, such as Google's Bidirectional Encoder Representations from Transformers (BERT) or OpenAI's Generative Pre-trained Transformer (GPT) series, are trained, bidirectionally or unidirectionally, on large volumes of generic textual data, designed to represent a wide array of common language use contexts and scenarios [16]. In specialized use contexts, these generic models often fail to accurately classify information because the language structures that require classification—their words, syntax, semantic context, and other textual or lexical signatures—are sparsely represented in the data that were used to train the generic model [17,18]. Some language models, such as ElutherAI's GPT-J-6B, are trained on open-source language modeling data sets curated from a mix of smaller open web crawl data sets alongside more technical papers from PubMedCentral and arXiv and can offer improved classification accuracy for technical applications [19]. Nevertheless, specialized tasks often require fine-tuning of general-purpose LLMs. Fine-tuning provides a way of overcoming the limitations

of generic LLMs by augmenting their training data with data selected to more accurately reflect the target domains toward which a model is fine-tuned. The fine-tuning process updates the model's parameters—the weights that affect which connections between the nodes and layers of a neural network become activated—and so helps a model permanently learn. Unlike practices, such as prompt engineering, that leave the underlying language model untouched, fine-tuning changes the model itself, yielding a new model optimized for the specific use case.

However, fine-tuning LLMs to perform technical, specialized tasks is expensive, because the target domain of a fine-tuned model is usually complex and technical—otherwise, fine-tuning would not be necessary—and it requires annotators with some degree of domain-level expertise, which comes with potentially significant financial and time costs. Indeed, one study of NER annotation speed found it can take between 10 and 30 seconds per sentence for experts to annotate named entities [8]. The gold-standard annotated BioSemantics corpus is composed of 163,219 sentences, which implies an optimal annotation time of over 11 weeks at 40 hours per week (453.39 h) [20]. This estimate, of course, excludes the time required for annotator training and interannotator reliability assessments, and because fine-tuning adjusts many or all of the model's parameters, it consumes computational resources. Time and power consumption for fine-tuning scales with training data size [21,22] and with the size of the underlying model that is computed. As of the date of writing, for example, it would be unrealistic to fine-tune very large models such as GPT-4.

These limitations notwithstanding, it is increasingly recognized that long-standing presumptions about sufficiently large training data sets are likely substantially inflated [23]. We suspect this comes from a research and development environment dominated by a significant focus on promulgating new models that can claim to be state-of-the-art (SOTA) based on some preidentified benchmark. In a research environment dominated by so-called "SOTA chasing," ever larger data sets are often required to eke out minor performance improvements over the previous benchmarks. Notably, development teams from disciplines with generally small research budgets have found that fine-tuning can result in substantial performance improvements from relatively small amounts of expert-annotated data [13,24] or from a combination of prelearning and transfer learning followed by a brief fine-tuning phase [25]. In one case, significant improvements over the baseline were derived from training samples as small as 50 lemmas [13]. Despite the growing recognition that smaller gold-standard training sets can provide

substantial performance improvements, there is little in the way of actionable guidance for sample size and sample curation.

The primary goal of this study is to establish some initial baselines for sample size considerations in terms of training set size and relevant entity density for NER applications in specialized technical domains. To that end, we have conducted a fine-tuning experiment that compares the performance improvements resulting from 2500 randomly selected training data sets stratified by size. These training sets were used to fine-tune 4 distinct language models to perform NER in a highly specific language domain: the identification of 2 internal components (conflict sources and conflict targets) in conflicts of interest (COI) disclosures. The results presented below indicate that only relatively small samples are required for substantial improvement. They also demonstrate a rapidly diminishing marginal return for larger sample sizes. In other words, while larger and larger sample sizes may be useful for “SOTA chasing,” their value for fine-tuning LLMs shrinks beyond a certain threshold, which we estimate below. These findings provide actionable guidance about how to select and generate fine-tuning samples by attending to issues of relevant token density. As such, they should have great value for NER applications that rely on them.

Literature Review

During our initial review of the literature, we were unable to locate any widely accepted, evidence-based guidance on appropriate sample sizes for training data in NER fine-tuning experiments. Therefore, to evaluate the state of the field, we conducted a literature search focused on identifying existing practices. We searched PubMed for prior relevant work to determine current sample size conventions in NER fine-tuning. We used a simple search strategy (“named entity recognition” OR “entity extraction”) AND (fine tuning OR transfer learning) AND (annotat*),” which returned 138 relevant papers. We reviewed each of these papers and extracted information related to human-annotated NER training sets. Specifically, for each paper, we assessed if a human-annotated training set was used, and if so, we extracted data on sample units, sample size, and any available sample size justification. In cases where authors described the size of human-annotated training sets on multiple levels (eg, number of documents, number of sentences, and number of entities), we prioritized units that would most effectively guide prospective sampling. This emphasis meant that we prioritized sentences (as they are comparable across document types and identifiable without annotation) over documents (which vary widely in length) or entities (which cannot be assessed until after annotation). In cases where multiple human-annotated samples were used, we noted the largest reported sample as indicative of the researchers’ sense of the sample necessary to conduct the research in its entirety.

Additionally, for each paper that made use of a human-annotated training set, we sought to identify any possible justifications for the chosen sample size. We anticipated that common justifications might include (1) collecting a sample sufficient to achieve target performance, (2) collecting a sample consistent with or larger than prior work, or (3) collecting a sample appropriate given relevant power calculations.

Of the papers surveyed, the majority (93/138, 67.4%) reported the use of human-annotated NER training data. The remaining (45/138, 32.6%) papers used only computational approaches to curate training data sets. Notably, many papers reported using a mix of human-annotated and computationally-annotated training sets or performing multiple experiments with different training sets. As long as any given paper used at least 1 human-annotated training set, it was included in the tally. Reported sample units varied quite widely across papers with many reporting only the number of documents used. Document types were similarly variable and specific to research contexts. For example, several papers reported training sample sizes as the number of clinical notes, number of published abstracts, or number of scraped tweets. In contrast, some papers reported sample size using non-context-specific measures such as sentences, entities, or tokens. Given this variety, we classified sample units as belonging to 1 of 6 common categories: clinical notes or reports, sentences, abstracts or papers, entities, tokens, or others. The most commonly used sample unit was clinical notes or reports (34/93, 37%) followed by sentences and papers or abstracts (21/93, 23%). Sample size ranges also varied widely by unit type, as would be expected. The smallest clinical notes or reports sample used a scant 17 documents [26], but this was likely a larger sample than the smallest reported sentence sample size of 100 [27]. Among the papers reporting nondocument type specific sample units, human-annotated data sets ranged from 1840 tokens to 79,401 tokens (mean 42,121 tokens); from 100 entities to 39,876 entities (mean 15,957 entities); and from 100 sentences to 360,938 sentences (mean 26,678 sentences). Details on the sample size range by sample type are available in [Table 1](#). Complete details on each paper’s approach to sample size are available in [Multimedia Appendix 1](#).

Of the 93 papers that used human-annotated NER training data, only 3 (3%) papers provided an explicit justification for the chosen sample size. In each case, the justification for the sample size was based on a reference to prior relevant work and determined to be as large or larger than a sample used in the previously published work [28-30]. Ultimately, the wide range of sample reporting practices and the broad lack of attention to sample size justification indicate a strong need for explicit sample selection guidance for fine-tuning NER models. This paper contributes to addressing this need.

Table 1. Unit types, number of papers by type, and sample size means and ranges.

Unit type	Papers (n=93), n (%)	Sample size, mean	Sample size, range
Clinical notes or reports	34 (37)	709	17-5098
Abstracts or papers	21 (23)	1966	20-7000
Sentences	21 (23)	26,678	100-360,938
Other	9 (10)	5979	47-25,678
Entities	5 (5)	15,957	100-39,876
Tokens	3 (3)	42,121	1840-79,401

Methods

Overview

The primary aim of this study was to evaluate sample size considerations for fine-tuning LLMs for domain- and context-specific NER tasks. Specifically, the goal was to evaluate how changes in retraining data set sizes and token density impact overall NER performance. To accomplish this task, we used stratified random samples of training sets to create 2500 fine-tuned instances of RoBERTa_base, GatorTron_base, RoBERTa_large, and GPT-2_large. In what follows, we describe (1) the data and target NER task, (2) the gold-standard annotation protocol, (3) the fine-tuning approach, and (4) our sample feature analysis.

Data Description and Context

We selected COI disclosures in biomedical literature as a highly domain-specific, technical language context suitable for the goals of this paper. In recent years, significant research efforts have been devoted to studying the effects of financial COI on the biomedical research enterprise [31-33], finding that COI is associated with favorable findings for sponsors [31], increased rates of “spin” in published reports [34], increased likelihood of trial discontinuation or nonpublication [35], editorial and peer reviewer biases [36], and increased adverse events rates for developed products [37]. Unfortunately, as compelling as this body of evidence is, a recent methodological review of research in this area indicates that most studies treat COI as a binary variable (present or absent) rather than quantifying COI rates or disaggregating COI types [32]. This limitation in the available evidence is, no doubt, driven in part by the data structures of COI reporting. When COI are reported, they are generally reported in unstructured or semistructured text. COI disclosure statements can also be quite long, as individual authors frequently receive and report multiple lines of funding from a wide variety of granting agencies and corporate sponsors. Ultimately, the lack of tabular data structures for COI makes it difficult to extract appropriate information [38] such as the sources and recipients of funding, the precise links between COI sources and recipients, or the quantity and degree of COI in a given disclosure statement.

These limitations notwithstanding, there has been some recent research leveraging informatics techniques, including NER, to transform text disclosure statements into tabular data [18,37]. Recently developed systems leverage NER to identify authors and sponsors as “PERSONs” and “ORGs,” respectively.

Secondary processing makes use of regular expressions to parse the types of relationships reported between each NER-identified PERSON and ORG. Since NER-tagging in this context is focused on identifying canonical entity types, applying these tools to COI disclosure statements may seem relatively straightforward at the outset. However, variances in reporting formats and the lack of specific training data on relevant entities present a number of challenges. In the first case, author identification is stymied by different journal guidelines for rendering author names. For example, a disclosure statement for Rudolf Virchow might be rendered as “Rudolf Virchow,” “Virchow,” “Dr. Virchow,” or “RLCV.” Likewise, pretrained NER models have not been found to offer high-quality, out-of-the-box performance for pharmaceutical company names [18]. Variations in incorporation type (Inc, LLC, GmbH, etc) typically induce entity boundary issues, and multinational companies often report national entity names (eg, Pfizer India), leading standard NER models to assign inappropriate geopolitical entity tags. Finally, effective NER on COI disclosure statements is also challenged by the atypical distribution of relevant tokens. It is not uncommon for a single sentence in a disclosure to have a dozen author names or a dozen company names, for example, when a disclosure statement lists all authors who have the same COI (eg, “such-and-such authors are employed at MSD”). These atypical sentence structures also occur when a single author has many COIs to disclose, as in, “RLCV receives consulting fees from MSD, Pfizer, GSK, Novartis, and Sanofi.”

To more clearly demonstrate these limitations, we provide the following authentic example from a COI disclosure statement published in a 2018 issue of the *World Journal of Gastrointestinal Oncology* [39]. The following shows the NER tagging performance of RoBERTa_base without fine-tuning:

Sunakawa Y[ORG] has received honoraria from Taiho Pharmaceutical[ORG], Chugai Pharma[ORG], Yakult Honsha[ORG], Takeda[ORG], Merck Serono[ORG], Bayer Yakuhin[ORG], Eli Lilly Japan[ORG], and Sanofi[ORG]; Satake H[ORG] has received honoraria from Bayer[ORG], Chugai Pharma[ORG], Eli Lilly Japan[ORG], Merck Serono[ORG], Takeda[ORG], Taiho Pharmaceutical[ORG] and Yakult Honsha[ORG]; Ichikawa W[ORG] has received honoraria from Chugai Pharma[ORG], Merck Serono[ORG], Takeda Pharmaceutical[ORG], and Taiho Pharmaceutical[ORG]; research funding from Chugai

Pharma[ORG], Takeda Pharmaceutical[ORG], and Taiho Pharmaceutical[ORG].

Furthermore, the following shows the NER tags provided by the human annotation team:

Sunakawa Y[PERSON] has received honoraria from Taiho Pharmaceutical[ORG], Chugai Pharma[ORG], Yakult Honsha[ORG], Takeda[ORG], Merck Serono[ORG], Bayer Yakuhin[ORG], Eli Lilly Japan[ORG], and Sanofi[ORG]; Satake H[PERSON] has received honoraria from Bayer[ORG], Chugai Pharma[ORG], Eli Lilly Japan[ORG], Merck Serono[ORG], Takeda[ORG], Taiho Pharmaceutical[ORG] and Yakult Honsha[ORG]; Ichikawa W[PERSON] has received honoraria from Chugai Pharma [ORG], Merck Serono[ORG], Takeda Pharmaceutical[ORG], and Taiho Pharmaceutical[ORG]; research funding from Chugai Pharma[ORG], Takeda Pharmaceutical[ORG], and Taiho Pharmaceutical[ORG].

It is evident that the base LLM classifier makes critical errors that make mapping COI relationships between researchers, funding streams, and funding sources impossible. In the above example, a base-trained classifier mistakenly tags PERSONs as ORGs; elsewhere, we have seen the opposite, where non-fine-tuned classifiers mistakenly identify companies, such as Novartis or Eli Lilly, as PERSONs. General purpose language models (such as BERT and GPT-3) are not well-suited to the NER task of classifying and linking named authors and disclosed payors (pharmaceutical companies, nonprofit foundations, federal funders, etc) because of challenges that arise from the aforementioned lack of standardized disclosure conventions for author names. Likewise, another challenge arises because these models are not well-trained on biomedical companies, nonprofit entities, and federal funders. In this study, as well as earlier research, we found that pharmaceutical companies—frequently named after founding families—are often tagged as PERSONs rather than ORGs. Finally, the linguistic signature of COI disclosure statements is distinctive: COI statements deploy semicolons in nonstandard ways. For large research teams, a single disclosure sentence can cover the length of a long paragraph, and grammatical conventions that govern the relationship between subjects, direct objects, and indirect objects are often elided or circumvented in favor of brevity, which makes linking authors to payors and payors to type of payment challenging. At the same time, the linguistic conventions used for disclosure statements vary between and even within journals, rendering rule-based NER approaches unfeasible. As such, the task of identifying and linking authors to payors and payment types in COI statements is an ideal use case for fine-tuning parameter-dense language models based on gold-standard human annotated COI statements.

Data Sources and Preprocessing

The data used for fine-tuning COI-relevant NER tags in this study come from COI disclosure statements drawn from 490 papers published in a diverse range of biomedical journals. The selected disclosure statements were randomly sampled from a preexisting data set of 15,374 statements with artificial

intelligence-identified COI [40]. The original data set was created by extracting all PubMed-indexed COI statements in 2018. At the time of download, there were 274,246 papers with a COI-statement field in the PubMed XML file. The substantial majority of these are statements of no conflict disclosure, and thus collected statements were analyzed using a custom machine learning-enhanced NER system that can reliably identify relationships between funding entities and named authors [18,37]. The sample used in this study was drawn from the population of COI statements with artificial intelligence-confirmed conflict disclosures.

Two annotators independently tagged named entities in the collected COI statements as either people (PERSON) or organizations (ORG). The PERSON tag was applied to all named authors, regardless of the format of the name. This included initials with and without punctuation, for example, “JAD” or “J.A.D” as well as full names “Jane A. Doe” or names with titles “Dr. Doe.” ORG tags were applied to named pharmaceutical companies, nonprofit organizations, and funding agencies. To ensure that NER tagging was consistent, a random sample of 200 COI statements was tagged by both annotators and assessed for interannotator agreement using interclass correlation coefficient for unit boundaries and Cohen κ for entity type agreement. The raters had 98.3% agreement on unit boundaries (interclass correlation coefficient=0.87, 95% CI 0.864-0.876). For named entities with identical unit boundaries, the classification (PERSON or ORG) agreement was 99.6% (κ =0.989). After this high degree of interrater reliability was established, the annotators independently annotated the remaining COI statements. Prior to training the language model, a third rater reconciled the few annotation disagreements in the initial interrater reliability sample.

Model Fine-Tuning and Analysis

A subset (147/490, 30%) of the annotated disclosure statements was reserved to serve as an evaluation set. The remaining 343 statements were used to generate 2500 training sets for subsequent experimentation. Each set was created by randomly selecting an N size in 5 preidentified strata of 40 possible sample sizes, at the statement level. The strata included size ranges of 1-40, 41-80, 81-120, 121-160, and 161-200. Once each N size was selected, a random sample of COI statements at that N size was derived. We created 500 random samples within each stratum.

We fine-tuned 4 commonly used language models using the open-source *spaCy* NLP library (version 3.2.1, running on Python version 3.9.7). To ensure the repeatability of results and to make the fine-tuning process as accessible as possible to research teams, we used *spaCy*'s default configuration settings for NER. The selected models included RoBERTa_base, GatorTron_base, RoBERTa_large, and GPT-2_large; for the latter 3, we used the *spacy-transformers* package to access these models through Hugging Face's *transformers* library. These models were selected to provide a range of parameter sizes (125M to 744M) and to allow for a comparison between language models trained on general use, as well as on biomedical texts specifically. Fine-tuning was performed on *spaCy*'s pretrained transformer pipeline, with only the *transformer* and

NER pipeline components enabled in the configuration file. All fine-tuning processes were run on a high-performance computing cluster at North Dakota State University's Center for Computationally Assisted Science and Technology, using AMD EPYC central processing units and NVIDIA graphics processing units. Preprocessing and tokenization were done using *spaCy*'s built-in tokenizer; training runs were optimized with the Adam algorithm, with decay rates of 0.9 (beta1) and 0.999 (beta2) and a learning rate of 0.01. For each training run, *spaCy* was set to check NER classifications against the test set after every 200 iterations within an epoch, to generate language models at regular intervals during the training process, and to stop whenever additional training steps failed to improve the classification metrics. We then extracted the highest-scoring language model from each set, for a total of 2500 fine-tuned language models.

Each of the 2500 retraining sets was subsequently categorized by sample size (measured in the number of sentences) and relevant entity density (entities per sentence [EPS]). Sentence boundaries were determined using the sentencizer in the *R tidytext* (0.3.4) library [41]. Sentences were used to provide a more regularized comparator as disclosure statements vary widely in length. We also focus on sentences as opposed to tokens since the number of sentences in a sample can be identified prospectively (ie, prior to annotation). Multiple regression was used to assess the linear relationship between sample size (number of sentences), entity density (EPS), and trained model F_1 -score. Additionally, we used single-predictor threshold regression models for the number of sentences and EPS to evaluate the possibility of diminishing marginal returns from increased sample size or taken density [42]. Threshold regression offers an effective way to model and evaluate nonlinear relationships, and as the term suggests, to identify any threshold effects. Multiple threshold models are available, and our approach relies on a hinge model that can be expressed as follows:



All statistical tests were performed in R (version 4.2.2; The R Foundation) and the threshold modeling was performed using the R *chnsgpt* package [43].

Ethical Considerations

This study does not include human subjects research (no human subjects experimentation or intervention was conducted) and so does not require institutional review board approval.

Results

The 2500 sets ranged from 1 to 200 disclosure statements with an average of 100 (SD 57.42). The number of sentences in each fine-tuning set ranged from 5 to 1031, with an average of 525.2 (SD 294.13). The tagged entity density ranged from 0.771 to 1.72 EPS, with an average of 1.34 (SD 0.14). Fine-tuned model

performance on NER tasks ranged from F_1 -score=0.3 to F_1 -score=0.96. The top F_1 -score for each architecture was 0.72 for GPT-2_large, 0.92 for GatorTron_base, 0.94 for RoBERTa_base, and 0.96 for RoBERTa_large. Data set and model descriptive statistics are available in Table 2.

Multiple linear regressions were used to assess and compare the relationship between the independent variables (number of sentences and EPS) and the overall model performances (measured by F_1 -score) for each architecture. EPS and number of sentences predictors correlate weakly (Pearson r =0.28, P <.001), and diagnostic tests for multicollinearity indicate that the variables do not violate the Klein rule of thumb and have a low variance inflation score (1.11) and high tolerance (0.9) [44].

All models were statistically significant with multiple R^2 ranging from 0.6057 to 0.7896 (all P <.001). EPS and the number of sentences were significant predictors of F_1 -scores in all cases (P <.001), except for the GPT-2_large model, where EPS was not a significant predictor (P =.184). Standardized regression coefficients and full model results are available in Table 3.

This study focuses primarily on total sentences as our measure of data size. This is because the number of sentences can be identified prospectively (prior to annotation) and is comparable across data sets with different document lengths. However, it should be noted that other measures of sample size are similarly predictive of F_1 -scores. The total number of relevant entities per training data set correlates very closely with the number of sentences (Pearson r =0.998, P <.001). This high collinearity makes it inadvisable to fit regression models with both predictors. We did, however, fit a series of models with EPS and a number of relevant entities as predictors. In all cases, the results were quite similar to those reported in Table 3. Specific values are available in Multimedia Appendix 2. It is notable that, in all cases, the multiple R^2 for models with EPS and the number of relevant entities as predictors are lower than the counterpart models with EPS and number of sentences. Subsequent pairwise ANOVA, however, indicates that there are no significant differences in model fit. ANOVA P values were 0.85 for RoBERTa_base, 0.74 for GatorTron_base, 0.93 for RoBERTa_large, and 0.53 for GPT-2_large.

Threshold regression models were also used to assess the possibility of diminishing marginal returns on training data sizes and EPS for each model and model architecture. All threshold models indicate that there was a diminishing marginal return from increased training data set sample size measured by number of sentences. Point estimates ranged from 439 for RoBERTa_large to 527 for GPT-2_large. Likewise, the threshold models indicate a diminishing marginal return for EPS with point estimates between 1.36 and 1.38. Complete threshold regression results are available in Table 4. Single predictor plots are available in Figure 1, with technical threshold model plots shown in Multimedia Appendix 2.

Table 2. Descriptive statistics of training sets and model performance.

Descriptive statistics	Value, range	Value, mean (SD)
Number of disclosure statements	1-200	100.0 (57.42)
Number of tokens	4-1402	712.9 (405.94)
Number of sentences	5-1031	525.2 (294.13)
Entities per sentence	0.771-1.72	1.34 (0.14)
RoBERTa_base F_1 -score	0.43-0.94	0.81(0.13)
GatorTron_base F_1 -score	0.37-0.92	0.84 (0.13)
RoBERTa_large F_1 -score	0.44-0.96	0.84 (0.14)
GPT-2_large F_1 -score	0.30-0.72	0.58 (0.12)

Table 3. Standardized multiple linear regression results by architecture.

Model (parameters)	β_{EPS}^a	β_{sent}	F test (df)	P value ^b	Multiple R^2
RoBERTa_base (125M)	0.04 ^c	0.78 ^c	2034 (22, 497)	<.001	0.6197
GatorTron_base (345M)	0.05 ^c	0.79 ^c	2236 (22, 497)	<.001	0.6417
RoBERTa_large (355M)	0.05 ^c	0.76 ^c	1918 (22, 497)	<.001	0.6057
GPT-2_large (774M)	-0.01	0.89 ^c	4685 (22, 497)	<.001	0.7896

^aEPS: entities per sentence.

^bIndividual predictor P values for Beta_sent were <.001 for all models. P values for Beta_EPS were <.001 in all cases except for the GPT-2_large model where EPS was not a significant predictor ($P=.184$)

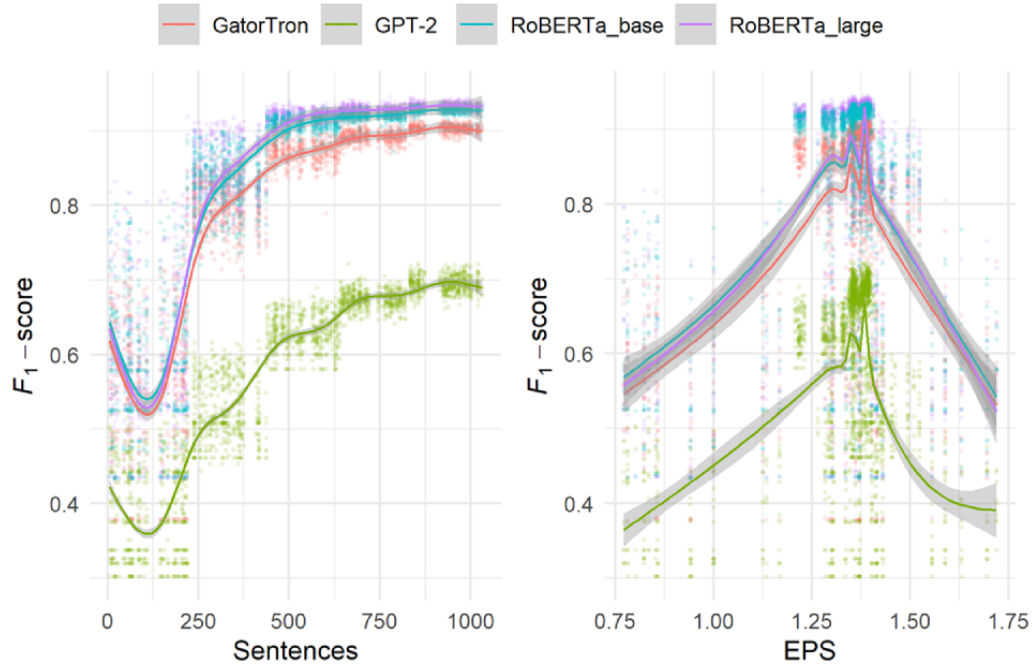
^cPredictor results are significant at the $P<.01$ level.

Table 4. Threshold regression point estimates and 95% confidence intervals for number of sentences and EPS^a by architecture.

Model (parameters)	Number of sent threshold, estimate (95% CI)	EPS threshold, estimate (95% CI)
RoBERTa_base (125M)	448 (437-456)	1.36 (1.35-1.37)
GatorTron_base (345M)	448 (409-456)	1.36 (1.36-1.38)
RoBERTa_large (355M)	439 (409-451)	1.36 (1.35-1.38)
GPT-2_large (774M)	527 (511-540)	1.38 (1.36-1.38)

^aEPS: entities per sentence.

Figure 1. Single predictor plots for the number of sentence (left) and EPS (right). Fit with a generalized additive model. EPS: entities per sentence.



Discussion

Principal Findings

Our review of the available literature on human-annotated training data for NER fine-tuning indicates that there is a strong need for useful guidance on requisite sample sizes. Reported sample units and sizes vary widely, providing little foundation for prospective approaches to sample curation. Given the significant time and costs associated with gold-standard annotation, it is critical that researchers and practitioners can effectively determine appropriate samples before fine-tuning neural network language models. The results of the experiment presented here provide initial actionable guidance for the development of gold-standard annotated training sets for NER fine-tuning in highly specific, specialized domains. Specifically, they indicate that contrary to common assumptions, transformer-based language models can be optimized for new tasks using relatively small amounts of training data. Furthermore, the results presented here indicate that NER fine-tuning is subject to threshold effects whereby there are diminishing marginal returns from increased sample sizes. Our data revealed that a scant 439 sentences were sufficient to reach that threshold with RoBERTa_large. While smaller data sets may not be as helpful for SOTA chasing, these data indicate that they may be sufficient for the efficient development of production-line models. These findings are consistent with the growing multidisciplinary body of literature demonstrating the efficacy of smaller sample sizes for fine-tuning [13,23,24]. Additionally, we note that given prior estimates for NER annotation rates, a sample of approximately 450 sentences would take between 74 and 225 minutes to annotate [8].

Importantly, the data provided here also indicate that neither model size nor content area-specific foundational training data may be essential for maximizing performance, but that model architecture is. RoBERTa_base, GatorTron_base, and

RoBERTa_large all achieved comparable performance levels in terms of maximum F_1 -score with similarly low training sample sizes. GPT-2_large, despite being the largest model tested, showed the worst performance on our NER tasks. On the one hand, neither finding is surprising. The foundational paper by Devlin et al [16] on the BERT transformer architecture suggests that BERT's capacity for fine-tuning for NLP tasks, such as classification, is better compared with GPT-based models, and a recent Microsoft Research paper argues that general-language models, such as GPT-4, can perform as well or better on domain-specific language tasks—specifically as they relate to medicine—than models trained on language specific to that domain [45]. But where the latter study focused on a very LLM built with reinforcement learning from human feedback and designed to be responsive to prompting, we found that for smaller—and therefore more tunable—models, fine-tuning with domain-specific texts yields significant performance improvements. For domain-specific NER tasks, then, architecture differences may matter most: decoder-based unidirectional architectures may be better suited for sentence generation, while encoder- or decoder-based bidirectional architectures better capture sentence-level contexts that are essential to NER tasks.

The results presented here also indicate that there are similar threshold effects for token density. That is, selecting or synthetically creating specifically token-rich samples may not improve model performance. Unlike the sample size data that indicate a diminishing marginal return, the hinge model for token density shows a substantial decrease in overall performance after the EPS threshold is achieved. We note that these threshold point estimates and narrow 95% CIs converge on the average EPS (1.34) of the 2500 training sets, and this suggests that the relevant entity density of training data needs to approximate the relevant entity density of testing and production-line data.

This finding is especially relevant given the increasing interest in artificial training data generated by LLMs. While the insights presented here indicate that fine-tuning training data can be much smaller than generally anticipated, high-quality small training data sets still require adequate funding and time to pay, train, and deploy human annotators. In response, some research seeks to leverage LLMs as sources of training data for subsequent fine-tuning of smaller neural network models [46]. This is an intriguing line of research worthy of further scrutiny. However, it is notable that our findings about relevant token density suggest that artificially generated data must mirror real data in terms of token density. If the token density is too low or too high, we can expect to see reduced model performance when compared with naturally derived training data and high-quality expert annotation.

While these findings provide an important initial foundation for fine-tuning sample size considerations in NER applications, the specifically identified thresholds may not apply to markedly different NER use cases. This study focused on fine-tuning PERSON and ORG tags, entity types that are well-represented across the heterogeneous data sources that are used to train LLMs. Bioinformatics use cases that focus on entity types that are more unique to biomedical contexts (eg, symptoms, chemicals, diseases, genes, and proteins) or that require generating new entity categories may require larger training samples to optimize LLM performance. Additionally, this study focuses on semistructured natural language (disclosure statements). While we would expect similar guidelines to apply for NER in other semistructured biomedical contexts (eg, research papers, clinical notes, abstracts, and figure or image annotations), the threshold guidance here may not apply well to less formalized linguistic contexts.

Conclusion

The emergence of LLMs offers significant potential for improving NLP applications in biomedical informatics, with research demonstrating the advantages of fine-tuned, domain-specific language models for health care applications

[47] and environmental costs [22]. However, given the novelty of these solutions, there is a general dearth of actionable guidelines on how to efficiently fine-tune language models. In the context of NER applications, this study demonstrates that there is a general lack of consensus and actionable guidance on sample size selection concerns for fine-tuning LLMs. Training sets reporting units and sample size varied widely in the published literature, with samples ranging from 100 sentences to 35,938 sentences for training sets. Additionally, human-annotated training set sample sizes are seldom justified or explained. In the rare cases where sample size is discussed explicitly, justifications focus narrowly on simple size comparisons to previously published efforts in a similar domain. In this context, biomedical informatics researchers could benefit from actionable guidelines about sample size considerations for fine-tuning LLMs.

The data presented here provide sample size guidance for fine-tuning LLMs drawn from an experiment on 2500 gold-standard human annotated fine-tuning samples. Specifically, the data demonstrate the importance of both sample sizes as measured in the number of sentences and relevant token density for training data curation. Furthermore, the findings indicate that both sample size and token density can be subject to threshold limitations where increased sample size or token density do not confer additional performance benefits. In this study, sample sizes of greater than 439-527 sentences failed to produce meaningful accuracy improvements. This suggests that researchers interested in leveraging LLMs for NER applications can save considerable time, effort, and funding, which has been historically devoted to producing gold-standard annotations. The data presented here also indicate that the relevant token density of training samples should reliably approximate the relevant token density of real-world cases. This finding has important ramifications for the production of synthetic data which may or may not effectively approximate real-world cases. The findings presented here can directly inform future research in health policy informatics and may also be applicable to a wider range of health and biomedical informatics tasks.

Acknowledgments

This work was supported by the National Institute of General Medical Sciences of the National Institutes of Health under award R01GM141476. The funder did not participate in the study design, conduct, or preparation of findings. This work used resources of the Center for Computationally Assisted Science and Technology at North Dakota State University, which were made possible in part by the National Science Foundation's Major Research Instrumentation Program (award 2019077).

Authors' Contributions

SSG and ZPM designed the study. ZPM implemented the fine-tuning pipelines. MSK and JTJ provided ground-truth annotations. JCSE and SNR conducted a review of prior findings. SSG conducted the statistical analyses. All authors participated in the interpretation of findings, drafting, and revision.

Conflicts of Interest

SSG reports grant funding from National Institute of General Medical Sciences (NIGMS) and the Texas Health and Human Services Commission. ZPM reports grant funding from NIGMS and National Science Foundation. SNR reports grant funding from the National Institute of Neurological Disorders and Stroke. JBB reports grant funding from NIGMS, National Science Foundation, and Blue Cross Blue Shield/Health Care Service Corporation. JRF reports grant funding from NIGMS, National Institute of Mental Health (NIMH), National Institute of Allergy and Infectious Diseases (NIAID), National Library of Medicine

(NLM), Health Care Cost Institute, Austin Public Health, Texas Child Mental Health Care Consortium, Texas Alzheimer Research and Care Consortium, and the Michael & Susan Dell Foundation. JFR also reports receiving a grant from the NIH Division of Loan Repayment. All other authors report no conflicts of interest.

Multimedia Appendix 1

Review of sample sizes and justifications.

[\[DOCX File, 131 KB - ai_v3i1e52095_app1.docx\]](#)

Multimedia Appendix 2

Detailed statistical results and threshold model plots.

[\[DOCX File, 133 KB - ai_v3i1e52095_app2.docx\]](#)

References

1. Nadkarni PM, Ohno-Machado L, Chapman WW. Natural language processing: an introduction. *J Am Med Inform Assoc* 2011;18(5):544-551 [FREE Full text] [doi: [10.1136/amiajnl-2011-000464](https://doi.org/10.1136/amiajnl-2011-000464)] [Medline: [21846786](https://pubmed.ncbi.nlm.nih.gov/21846786/)]
2. Aronson AR. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. *Proc AMIA Symp* 2001:17-21 [FREE Full text] [Medline: [11825149](https://pubmed.ncbi.nlm.nih.gov/11825149/)]
3. Yang X, Bian J, Fang R, Bjarnadottir RI, Hogan WR, Wu Y. Identifying relations of medications with adverse drug events using recurrent convolutional neural networks and gradient boosting. *J Am Med Inform Assoc* 2020;27(1):65-72 [FREE Full text] [doi: [10.1093/jamia/ocz144](https://doi.org/10.1093/jamia/ocz144)] [Medline: [31504605](https://pubmed.ncbi.nlm.nih.gov/31504605/)]
4. Ahmed A, Abbasi A, Eickhoff C. Benchmarking modern named entity recognition techniques for free-text health record deidentification. *AMIA Jt Summits Transl Sci Proc* 2021;2021:102-111 [FREE Full text] [Medline: [34457124](https://pubmed.ncbi.nlm.nih.gov/34457124/)]
5. Uzuner O, Solti I, Cadag E. Extracting medication information from clinical text. *J Am Med Inform Assoc* 2010;17(5):514-518 [FREE Full text] [doi: [10.1136/jamia.2010.003947](https://doi.org/10.1136/jamia.2010.003947)] [Medline: [20819854](https://pubmed.ncbi.nlm.nih.gov/20819854/)]
6. Alfattni G, Belousov M, Peek N, Nenadic G. Extracting drug names and associated attributes from discharge summaries: text mining study. *JMIR Med Inform* 2021;9(5):e24678 [FREE Full text] [doi: [10.2196/24678](https://doi.org/10.2196/24678)] [Medline: [33949962](https://pubmed.ncbi.nlm.nih.gov/33949962/)]
7. Pradhan S, Elhadad N, South BR, Martinez D, Christensen L, Vogel A, et al. Evaluating the state of the art in disorder recognition and normalization of the clinical narrative. *J Am Med Inform Assoc* 2015;22(1):143-154 [FREE Full text] [doi: [10.1136/amiajnl-2013-002544](https://doi.org/10.1136/amiajnl-2013-002544)] [Medline: [25147248](https://pubmed.ncbi.nlm.nih.gov/25147248/)]
8. Chen Y, Lask TA, Mei Q, Chen Q, Moon S, Wang J, et al. An active learning-enabled annotation system for clinical named entity recognition. *BMC Med Inform Decis Mak* 2017;17(Suppl 2):82 [FREE Full text] [doi: [10.1186/s12911-017-0466-9](https://doi.org/10.1186/s12911-017-0466-9)] [Medline: [28699546](https://pubmed.ncbi.nlm.nih.gov/28699546/)]
9. Lederman A, Lederman R, Verspoor K. Tasks as needs: reframing the paradigm of clinical natural language processing research for real-world decision support. *J Am Med Inform Assoc* 2022;29(10):1810-1817 [FREE Full text] [doi: [10.1093/jamia/ocac121](https://doi.org/10.1093/jamia/ocac121)] [Medline: [35848784](https://pubmed.ncbi.nlm.nih.gov/35848784/)]
10. Idnay B, Dreisbach C, Weng C, Schnall R. A systematic review on natural language processing systems for eligibility prescreening in clinical research. *J Am Med Inform Assoc* 2021;29(1):197-206 [FREE Full text] [doi: [10.1093/jamia/ocab228](https://doi.org/10.1093/jamia/ocab228)] [Medline: [34725689](https://pubmed.ncbi.nlm.nih.gov/34725689/)]
11. Kreimeyer K, Foster M, Pandey A, Arya N, Halford G, Jones SF, et al. Natural language processing systems for capturing and standardizing unstructured clinical information: a systematic review. *J Biomed Inform* 2017;73:14-29 [FREE Full text] [doi: [10.1016/j.jbi.2017.07.012](https://doi.org/10.1016/j.jbi.2017.07.012)] [Medline: [28729030](https://pubmed.ncbi.nlm.nih.gov/28729030/)]
12. Demner-Fushman D, Chapman WW, McDonald CJ. What can natural language processing do for clinical decision support? *J Biomed Inform* 2009;42(5):760-772 [FREE Full text] [doi: [10.1016/j.jbi.2009.08.007](https://doi.org/10.1016/j.jbi.2009.08.007)] [Medline: [19683066](https://pubmed.ncbi.nlm.nih.gov/19683066/)]
13. Manjavacas Arevalo E, Fonteyn L. Non-parametric word sense disambiguation for historical languages. : Association for Computational Linguistics; 2022 Presented at: Proceedings of the 2nd International Workshop on Natural Language Processing for Digital Humanities; November 20, 2022; Taipei, Taiwan p. 123-134 URL: <https://aclanthology.org/2022.nlp4dh-1.16>
14. Ziems C, Held W, Shaikh O, Chen J, Zhang Z, Yang D. Can large language models transform computational social science? arXiv Preprint posted online on April 12, 2023 [FREE Full text] [doi: [10.1162/coli_a_00502](https://doi.org/10.1162/coli_a_00502)]
15. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. *PLOS Digit Health* 2023;2(2):e0000198 [FREE Full text] [doi: [10.1371/journal.pdig.0000198](https://doi.org/10.1371/journal.pdig.0000198)] [Medline: [36812645](https://pubmed.ncbi.nlm.nih.gov/36812645/)]
16. Devlin J, Chang MW, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. arXiv Preprint posted online on October 11, 2018 [FREE Full text]
17. Liu X, Hersch GL, Khalil I, Devarakonda M. Clinical trial information extraction with BERT. : IEEE; 2021 Presented at: 2021 IEEE 9th International Conference on Healthcare Informatics (ICHI); August 09-12, 2021; Victoria, BC, Canada p. 505-506. [doi: [10.1109/ichi52183.2021.00092](https://doi.org/10.1109/ichi52183.2021.00092)]

18. Graham SS, Majdik ZP, Clark D, Kessler MM, Hooker TB. Relationships among commercial practices and author conflicts of interest in biomedical publishing. *PLoS One* 2020;15(7):e0236166 [FREE Full text] [doi: [10.1371/journal.pone.0236166](https://doi.org/10.1371/journal.pone.0236166)] [Medline: [32706798](https://pubmed.ncbi.nlm.nih.gov/32706798/)]
19. Gao L, Biderman S, Black S, Golding L, Hoppe T, Foster C, et al. The Pile: an 800GB dataset of diverse text for language modeling. *arXiv Preprint* posted online on December 31, 2020 [FREE Full text]
20. Akhondi SA, Klenner AG, Tyrchan C, Manchala AK, Boppana K, Lowe D, et al. Annotated chemical patent corpus: a gold standard for text mining. *PLoS One* 2014;9(9):e107477 [FREE Full text] [doi: [10.1371/journal.pone.0107477](https://doi.org/10.1371/journal.pone.0107477)] [Medline: [25268232](https://pubmed.ncbi.nlm.nih.gov/25268232/)]
21. Ciosici MR, Derczynski L. Training a T5 using lab-sized resources. *arXiv Preprint* posted online on August 25, 2022 [FREE Full text] [doi: [10.5260/chara.21.2.8](https://doi.org/10.5260/chara.21.2.8)]
22. Luccioni AS, Viguier S, Ligozat AL. Estimating the carbon footprint of BLOOM, a 176B parameter language model. *arXiv Preprint* posted online on November 3, 2022 [FREE Full text]
23. Widner K, Virmani S, Krause J, Nayar J, Tiwari R, Pedersen ER, et al. Lessons learned from translating AI from development to deployment in healthcare. *Nat Med* 2023;29(6):1304-1306. [doi: [10.1038/s41591-023-02293-9](https://doi.org/10.1038/s41591-023-02293-9)] [Medline: [37248297](https://pubmed.ncbi.nlm.nih.gov/37248297/)]
24. Majdik ZP, Wynn J. Building better machine learning models for rhetorical analyses: the use of rhetorical feature sets for training artificial neural network models. *Tech Commun Q* 2022;32(1):63-78. [doi: [10.1080/10572252.2022.2077452](https://doi.org/10.1080/10572252.2022.2077452)]
25. Weber L, Münchmeyer J, Rocktäschel T, Habibi M, Leser U. HUNER: improving biomedical NER with pretraining. *Bioinformatics* 2020;36(1):295-302 [FREE Full text] [doi: [10.1093/bioinformatics/btz528](https://doi.org/10.1093/bioinformatics/btz528)] [Medline: [31243432](https://pubmed.ncbi.nlm.nih.gov/31243432/)]
26. Doan S, Xu H. Recognizing medication related entities in hospital discharge summaries using support vector machine. *Proc Int Conf Comput Ling* 2010;2010:259-266 [FREE Full text] [Medline: [26848286](https://pubmed.ncbi.nlm.nih.gov/26848286/)]
27. Soysal E, Wang J, Jiang M, Wu Y, Pakhomov S, Liu H, et al. CLAMP—a toolkit for efficiently building customized clinical natural language processing pipelines. *J Am Med Inform Assoc* 2018;25(3):331-336 [FREE Full text] [doi: [10.1093/jamia/ocx132](https://doi.org/10.1093/jamia/ocx132)] [Medline: [29186491](https://pubmed.ncbi.nlm.nih.gov/29186491/)]
28. Furrer L, Jancso A, Colic N, Rinaldi F. OGER++: hybrid multi-type entity recognition. *J Cheminform* 2019;11(1):7 [FREE Full text] [doi: [10.1186/s13321-018-0326-3](https://doi.org/10.1186/s13321-018-0326-3)] [Medline: [30666476](https://pubmed.ncbi.nlm.nih.gov/30666476/)]
29. Zhan X, Humbert-Droz M, Mukherjee P, Gevaert O. Structuring clinical text with AI: old versus new natural language processing techniques evaluated on eight common cardiovascular diseases. *Patterns (N Y)* 2021;2(7):100289 [FREE Full text] [doi: [10.1016/j.patter.2021.100289](https://doi.org/10.1016/j.patter.2021.100289)] [Medline: [34286303](https://pubmed.ncbi.nlm.nih.gov/34286303/)]
30. Chun HW, Tsuruoka Y, Kim JD, Shiba R, Nagata N, Hishiki T, et al. Automatic recognition of topic-classified relations between prostate cancer and genes using MEDLINE abstracts. *BMC Bioinformatics* 2006;7(Suppl 3):S4 [FREE Full text] [doi: [10.1186/1471-2105-7-S3-S4](https://doi.org/10.1186/1471-2105-7-S3-S4)] [Medline: [17134477](https://pubmed.ncbi.nlm.nih.gov/17134477/)]
31. Lundh A, Lexchin J, Mintzes B, Schroll JB, Bero L. Industry sponsorship and research outcome. *Cochrane Database Syst Rev* 2017;2(2):MR000033 [FREE Full text] [doi: [10.1002/14651858.MR000033.pub3](https://doi.org/10.1002/14651858.MR000033.pub3)] [Medline: [28207928](https://pubmed.ncbi.nlm.nih.gov/28207928/)]
32. Graham SS, Karnes MS, Jensen JT, Sharma N, Barbour JB, Majdik ZP, et al. Evidence for stratified conflicts of interest policies in research contexts: a methodological review. *BMJ Open* 2022;12(9):e063501 [FREE Full text] [doi: [10.1136/bmjopen-2022-063501](https://doi.org/10.1136/bmjopen-2022-063501)] [Medline: [36123074](https://pubmed.ncbi.nlm.nih.gov/36123074/)]
33. Grundy Q, Dunn AG, Bourgeois FT, Coiera E, Bero L. Prevalence of disclosed conflicts of interest in biomedical research and associations with journal impact factors and altmetric scores. *JAMA* 2018;319(4):408-409 [FREE Full text] [doi: [10.1001/jama.2017.20738](https://doi.org/10.1001/jama.2017.20738)] [Medline: [29362787](https://pubmed.ncbi.nlm.nih.gov/29362787/)]
34. Lieb K, von der Osten-Sacken J, Stoffers-Winterling J, Reiss N, Barth J. Conflicts of interest and spin in reviews of psychological therapies: a systematic review. *BMJ Open* 2016;6(4):e010606 [FREE Full text] [doi: [10.1136/bmjopen-2015-010606](https://doi.org/10.1136/bmjopen-2015-010606)] [Medline: [27118287](https://pubmed.ncbi.nlm.nih.gov/27118287/)]
35. Roddick AJ, Chan FTS, Stefaniak JD, Zheng SL. Discontinuation and non-publication of clinical trials in cardiovascular medicine. *Int J Cardiol* 2017;244:309-315. [doi: [10.1016/j.ijcard.2017.06.020](https://doi.org/10.1016/j.ijcard.2017.06.020)] [Medline: [28622947](https://pubmed.ncbi.nlm.nih.gov/28622947/)]
36. van Lent M, Overbeke J, Out HJ. Role of editorial and peer review processes in publication bias: analysis of drug trials submitted to eight medical journals. *PLoS One* 2014;9(8):e104846 [FREE Full text] [doi: [10.1371/journal.pone.0104846](https://doi.org/10.1371/journal.pone.0104846)] [Medline: [25118182](https://pubmed.ncbi.nlm.nih.gov/25118182/)]
37. Graham SS, Majdik ZP, Barbour JB, Rousseau JF. Associations between aggregate NLP-extracted conflicts of interest and adverse events by drug product. *Stud Health Technol Inform* 2022;290:405-409 [FREE Full text] [doi: [10.3233/SHTI220106](https://doi.org/10.3233/SHTI220106)] [Medline: [35673045](https://pubmed.ncbi.nlm.nih.gov/35673045/)]
38. Grundy Q, Dunn AG, Bero L. Improving researchers' conflict of interest declarations. *BMJ* 2020;368:m422. [doi: [10.1136/bmj.m422](https://doi.org/10.1136/bmj.m422)] [Medline: [32161006](https://pubmed.ncbi.nlm.nih.gov/32161006/)]
39. Sunakawa Y, Satake H, Ichikawa W. Considering FOLFOXIRI plus bevacizumab for metastatic colorectal cancer with left-sided tumors. *World J Gastrointest Oncol* 2018;10(12):528-531 [FREE Full text] [doi: [10.4251/wjgo.v10.i12.528](https://doi.org/10.4251/wjgo.v10.i12.528)] [Medline: [30595807](https://pubmed.ncbi.nlm.nih.gov/30595807/)]
40. Graham SS, Majdik ZP, Clark D. Methods for extracting relational data from unstructured texts prior to network visualization in humanities research. *J Open Humanit Data* 2020;6(1):8 [FREE Full text] [doi: [10.5334/johd.21](https://doi.org/10.5334/johd.21)]
41. Silge J, Robinson D. tidytext: text mining and analysis using tidy data principles in R. *J Open Source Softw* 2016;1(3):37 [FREE Full text] [doi: [10.21105/joss.00037](https://doi.org/10.21105/joss.00037)]

42. Hastie TJ. Generalized additive models. In: Statistical models in S. Boca Raton, FL: CRC Press; 2017:249-307.
43. Fong Y, Huang Y, Gilbert PB, Permar SR. chngpt: threshold regression model estimation and inference. BMC Bioinformatics 2017;18(1):454 [FREE Full text] [doi: [10.1186/s12859-017-1863-x](https://doi.org/10.1186/s12859-017-1863-x)] [Medline: [29037149](https://pubmed.ncbi.nlm.nih.gov/29037149/)]
44. Ullah MI, Aslam M, Altaf S, Ahmed M. Some new diagnostics of multicollinearity in linear regression model. J Sains Malays 2019;48(9):2051-2060 [FREE Full text] [doi: [10.17576/jsm-2019-4809-26](https://doi.org/10.17576/jsm-2019-4809-26)]
45. Nori H, Lee YT, Zhang S, Carignan D, Edgar R, Fusi N, et al. Can generalist foundation models outcompete special-purpose tuning? case study in medicine. Microsoft. URL: <https://www.microsoft.com/en-us/research/publication/can-generalist-foundation-models-outcompete-special-purpose-tuning-case-study-in-medicine/> [accessed 2023-12-02]
46. Hsieh CY, Li CL, Yeh CK, Nakhost H, Fujii Y, Ratner A, et al. Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes. arXiv Preprint posted online on May 3, 2023 [FREE Full text] [doi: [10.18653/v1/2023.findings-acl.507](https://doi.org/10.18653/v1/2023.findings-acl.507)]
47. Yang R, Tan TF, Lu W, Thirunavukarasu AJ, Ting DSW, Liu N. Large language models in health care: development, applications, and challenges. Health Care Sci 2023;2(4):255-263 [FREE Full text] [doi: [10.1002/hcs2.61](https://doi.org/10.1002/hcs2.61)]

Abbreviations

BERT: Bidirectional Encoder Representations from Transformers
COI: conflicts of interest
EPS: entities per sentence
GPT: Generative Pre-trained Transformer
LLM: large language model
NER: named entity recognition
NLP: natural language processing
SOTA: state-of-the-art
USMLE: United States Medical Licensing Exam

Edited by K El Emam, B Malin; submitted 22.08.23; peer-reviewed by E Soysal, R Yang; comments to author 13.10.23; revised version received 13.12.23; accepted 30.03.24; published 16.05.24.

Please cite as:

*Majdik ZP, Graham SS, Shiva Edward JC, Rodriguez SN, Karnes MS, Jensen JT, Barbour JB, Rousseau JF
Sample Size Considerations for Fine-Tuning Large Language Models for Named Entity Recognition Tasks: Methodological Study
JMIR AI 2024;3:e52095*

URL: <https://ai.jmir.org/2024/1/e52095>

doi: [10.2196/52095](https://doi.org/10.2196/52095)

PMID: [38875593](https://pubmed.ncbi.nlm.nih.gov/38875593/)

©Zoltan P Majdik, S Scott Graham, Jade C Shiva Edward, Sabrina N Rodriguez, Martha S Karnes, Jared T Jensen, Joshua B Barbour, Justin F Rousseau. Originally published in JMIR AI (<https://ai.jmir.org>), 16.05.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR AI, is properly cited. The complete bibliographic information, a link to the original publication on <https://www.ai.jmir.org/>, as well as this copyright and license information must be included.

Original Paper

Augmenting Telepostpartum Care With Vision-Based Detection of Breastfeeding-Related Conditions: Algorithm Development and Validation

Jessica De Souza¹, MSc; Varun Kumar Viswanath¹, MSc; Jessica Maria Echterhoff², MSc; Kristina Chamberlain³, CNM, ARNP, IBCLC, MN; Edward Jay Wang¹, PhD

¹Department of Electrical and Computer Engineering, University of California, San Diego, La Jolla, CA, United States

²Department of Computer Science and Engineering, University of California, San Diego, La Jolla, CA, United States

³Division of Extended Studies, University of California, San Diego, La Jolla, CA, United States

Corresponding Author:

Jessica De Souza, MSc

Department of Electrical and Computer Engineering

University of California, San Diego

9500 Gilman Drive

La Jolla, CA, 92093

United States

Phone: 1 (858) 534 7013

Email: jdesouza@ucsd.edu

Abstract

Background: Breastfeeding benefits both the mother and infant and is a topic of attention in public health. After childbirth, untreated medical conditions or lack of support lead many mothers to discontinue breastfeeding. For instance, nipple damage and mastitis affect 80% and 20% of US mothers, respectively. Lactation consultants (LCs) help mothers with breastfeeding, providing in-person, remote, and hybrid lactation support. LCs guide, encourage, and find ways for mothers to have a better experience breastfeeding. Current telehealth services help mothers seek LCs for breastfeeding support, where images help them identify and address many issues. Due to the disproportional ratio of LCs and mothers in need, these professionals are often overloaded and burned out.

Objective: This study aims to investigate the effectiveness of 5 distinct convolutional neural networks in detecting healthy lactating breasts and 6 breastfeeding-related issues by only using red, green, and blue images. Our goal was to assess the applicability of this algorithm as an auxiliary resource for LCs to identify painful breast conditions quickly, better manage their patients through triage, respond promptly to patient needs, and enhance the overall experience and care for breastfeeding mothers.

Methods: We evaluated the potential for 5 classification models to detect breastfeeding-related conditions using 1078 breast and nipple images gathered from web-based and physical educational resources. We used the convolutional neural networks Resnet50, Visual Geometry Group model with 16 layers (VGG16), InceptionV3, EfficientNetV2, and DenseNet169 to classify the images across 7 classes: healthy, abscess, mastitis, nipple blebs, dermatosis, engorgement, and nipple damage by improper feeding or misuse of breast pumps. We also evaluated the models' ability to distinguish between healthy and unhealthy images. We present an analysis of the classification challenges, identifying image traits that may confound the detection model.

Results: The best model achieves an average area under the receiver operating characteristic curve of 0.93 for all conditions after data augmentation for multiclass classification. For binary classification, we achieved, with the best model, an average area under the curve of 0.96 for all conditions after data augmentation. Several factors contributed to the misclassification of images, including similar visual features in the conditions that precede other conditions (such as the mastitis spectrum disorder), partially covered breasts or nipples, and images depicting multiple conditions in the same breast.

Conclusions: This vision-based automated detection technique offers an opportunity to enhance postpartum care for mothers and can potentially help alleviate the workload of LCs by expediting decision-making processes.

(JMIR AI 2024;3:e54798) doi:[10.2196/54798](https://doi.org/10.2196/54798)

KEYWORDS

remote consultations; artificial intelligence; AI for health care; deep learning; detection model; breastfeeding; telehealth; perinatal health; image analysis; women's health; mobile phone

Introduction

Background

The benefits of breastfeeding for both the mother and baby, such as lower gastrointestinal infections in the child, more rapid maternal weight normalization after birth, and prolonged amenorrhea for the mother, are just a few examples of why physicians recommend breastfeeding for at least 6 months [1-5]. Breastfeeding rates are on the rise in the United States, with 83.2% of newborn infants being breastfed in 2019, thanks to increased education and promotion of its benefits [6]. Despite the compelling evidence, many families struggle to continue breastfeeding. Although 95% of mothers initiate breastfeeding, the continuation rate drops to <41% and <19% for exclusive breastfeeding at 3 and 6 months, respectively [7]. Parents who breastfeed may face issues, such as low milk supply, fatigue, medical problems, difficulties with feeding techniques or pain, and lack of social support [8-10].

Lactation consultant (LC) professionals specialize in breastfeeding, milk supply, breast and nipple issues, breast milk management, and prenatal education. LCs ensure a mother's smooth and painless transition into breastfeeding and increase the possibility of continued breastfeeding through 6 months or longer [11,12]. The availability of international board-certified LCs (IBCLCs) globally is limited. In 2021, there were 3.6 million births in the United States and only 18,500 LCs with IBCLC certification, a rate of 194 babies per LC a year. In low- and middle-income countries such as Brazil, for instance, there were 2.6 million births in the same year but only 154 certified LCs, resulting in a rate of 16,883 babies per LC per year. The high demand for LCs, coupled with geographic and financial barriers, underscores the need for better tools to improve access to specialized lactation services, especially in less urbanized areas where such resources are scarce, leading to decreased breastfeeding support [13-20].

Another issue is professional availability itself, as LCs often combine their practice with midwife nursing, splitting their time between prenatal visits, attending births, lactation consultations, and managing their patients, which can lead to professional exhaustion, burnout, and emotional stress [21-23]. Moreover, the predominantly independent practice of LCs outside the United States, without the support of clinics with sophisticated patient management and triage systems, further complicates their time management and patient organization [22,24].

Supporting LCs Through Tele-Lactation Services

Tele-lactation services facilitate text, audio, and video communication. This enables LCs to consult with patients from any location, reduces travel time, helps balance their workload, increases their availability to receive new patients, and provides quicker responses to their patients [20]. Complementing tele-lactation services, patient triaging using information systems allow LCs to prioritize in-person visits for severe cases requiring

physical assessment, while less critical cases can be handled remotely [25,26]. Prior research suggests that LCs would benefit from time-saving tools for efficient patient information delivery while focusing on mitigating prolonged interactions, helping alleviate the burden on these professionals with a load of patients [22,27]. As LCs often follow up with their patients up to weeks after birth to ensure positive breastfeeding outcomes, an easy-to-access system to monitor patient progress is essential for effective patient triage, facilitating consultation scheduling, holding remote consultations, or providing reassurance. However, LCs' current access to remote consultation systems lacks patient triaging tools and is not time efficient, indicating an area in need of development.

Our work proposes a novel method for the identification of breastfeeding-related conditions using convolutional neural networks (CNNs). We evaluated a self-curated data set containing 7 different breastfeeding conditions on 5 distinct CNN models. The assessment of breast conditions is vital as pain and discomfort experienced during breastfeeding is a major barrier faced by parents who want to continue breastfeeding their child. About 80% of mothers are estimated to experience nipple pain and fissures, while 20% are estimated to experience mastitis [28,29]. Our pipeline incorporates automatic detection of visually discernible painful breastfeeding-related conditions, such as nipple cracks and fissures related to poor latching and positioning; skin conditions, such as dermatitis, eczema, thrush, or herpes; and risk of mastitis spectrum issues, such as engorgement, abscess, and nipple blebs. The CNN model is used for automatic detection of breast conditions, which can benefit the triaging of remote lactation patients for faster and more efficient patient response based on their conditions.

Our work evaluated 5 distinct CNN models' ability to differentiate between healthy and various unhealthy breast conditions (including breast abscesses, dermatoses, engorgement, mastitis, nipple blebs, and nipple damage) by performing both multiclass and binary evaluations on 1078 breast images. We evaluated the model's performance using the data set with and without data augmentation techniques. The data were divided into training, validation, and testing sets, using k-fold cross-validation for robustness. Performance evaluation on the best model includes an average area under the curve (AUC) of 0.93 for all conditions after data augmentation and precise detection of healthy breasts (precision of 84.4%) and unhealthy breasts (average precision of 66%, SD 12.8%) for 6 conditions. For binary classification, we achieved, with the best model, an average AUC of 0.96 for all conditions after data augmentation and precise detection of healthy breasts (precision of 93.8%) and unhealthy breasts (precision of 83.5%). The breast images have been curated from perinatal education resources such as images and video recordings under various lighting, environments, and image-taking conditions, where we examined potential issues around how the images are taken and their impacts on performance. Finally, we provide insights into

future designs of user interfaces and guidance needed for the proper application of the system.

Related Work

Lactating Care Pipeline: In-Person, Remote, and Hybrid

Health care providers introduce breastfeeding options to expectant mothers, including educational materials in print or web-based, during prenatal care. The initiation of breastfeeding after delivery is timed according to the type of birth. Many hospitals worldwide follow the United Nations Children's Fund and World Health Organization baby-friendly initiative, prioritizing maternal and infant health and supporting mothers facing challenges [30,31]. After a child's birth, families often seek breastfeeding support from LCs, who typically offer hands-on consultations from birth until support is no longer required [18]. They conduct visual and physical evaluations of both mother and baby, assessing the baby's internal mouth structure, breast and nipple anatomy, and milk supply and ensuring proper attachment or repositioning of the baby to prevent nipple fissures. LCs may also introduce laser therapy as a treatment option for damaged nipples from breast pump misuse or issues with baby attachment [8]. The immersive approach of LCs is crucial for providing personalized and effective lactation support to mothers and infants.

Remote Lactation Care

The widespread adoption of smartphone communication apps, particularly WhatsApp (Meta Platforms, Inc), has transformed public health facilities, including family clinics in limited-income countries, offering various patient services such as appointment scheduling, health guidance, and vaccine campaign notifications [32-34]. WhatsApp has become a popular communication tool between LCs and patients, facilitating breastfeeding education and family support during the neonatal period [35,36]. During the COVID-19 pandemic, LCs transitioned to telehealth consultations using established smartphone apps such as WhatsApp, Instagram (Meta Platforms, Inc), and Facebook (Meta Platforms, Inc). LCs adapted their approach to maintain quality care despite resource limitations in remote consultations [37,38]. Similar to other practices requiring physical evaluation, LCs reimagined their methods when shifting from in-person to remote consultations, using communication and social media apps to reach and educate parents while having broader visibility in their community [37,39].

Remote lactation care presents challenges, including limited visibility during video calls, communication difficulties, and technical issues [18,40,41]. Despite challenges, remote care offers benefits, reducing the mother's sense of isolation, enabling faster feedback, and promoting effective communication and patient engagement for improved independent learning [17,18,22]. These benefits positively impact mothers' intentions in exclusive breastfeeding for up to 6 months and reduce the risk of breastfeeding cessation at 3 months by 25% [42].

Hybrid Lactation Care

Previous research showed that fully remote consultations work well for cases where geographic distance, transportation issues, or patient disease prevent in-person meetings between patients and providers. LCs often conduct remote consultations from their workplaces, including personal offices, clinics, or hospitals, especially when they are also midwives with on-call responsibilities [37]. They provide consultations for patients before birth, after birth, and in emergency cases where the mother is facing breastfeeding challenges [22]. Depending on the nature of the consultation, in-person or remote visits are chosen to meet the patient's specific needs. In summary, remote care complements in-person care, being a valuable resource for mothers seeking guidance, reassurance, and confidence, particularly in the absence of a supportive home environment [38].

LCs, especially those who are also midwives, have limited time availability due to demanding schedules and receiving numerous remote messages from patients daily, some requiring higher priority attention [22,43]. Manually sorting through patient messages to determine priority can be time consuming and inconvenient for mothers with urgent needs. Our work proposes a computer vision-based system to triage breast conditions, facilitating telehealth and assisting LCs in identifying patients who require immediate responses in remote settings.

Issues Associated With Breastfeeding

Breastfeeding pain is one of the reasons associated with breastfeeding cessation, which can be caused by issues such as poor attachment of the baby onto the breast, physical conditions of the mother or baby, misuse of breast pumps, oversupply of breast milk, and even environmental conditions [44]. These issues, if left untreated in the first few days after birth, can persist for weeks and pose a threat to breastfeeding continuity beyond 6 months. Some conditions can be fully mitigated when the mother receives orientation and education on the topic. In contrast, other conditions can be alleviated and managed for a better experience for the mother in the case of physical conditions, including nipple physiology, baby tongue-tie, jaw clenching, and excessive milk supply [28,45].

This study concentrates on conditions leading to breastfeeding pain and potential interruption. The first condition is the mastitis spectrum disorder, where about 20% of mothers who breastfeed may face it during their time breastfeeding. This disorder starts with the overproduction of milk and breast engorgement, which can cause milk passage obstruction in the form of galactoceles and nipple blebs. When not properly treated, a case of milk bleb or galactocele can evolve into phlegmon, bacterial, or inflammatory mastitis, which may require patients to treat it with medications and sometimes medical procedures to drain the inflammation fluids from the breast in case it becomes an abscess [46,47]. Conditions associated with mastitis are painful and include symptoms such as redness in the breast, influenza-like symptoms, hardened skin surface in the location of the milk blockage, formation of blisters in the nipple, and even blood in the milk [29,48].

The second condition is nipple damage caused by improper latching and positioning from the infant, excessive pressure from breast pumping devices, infant tongue-tie or palate abnormality, infant's arrhythmic milk expression, and even infant biting or jaw clenching [9,44]. Considering the cause of nipple damage, 80% of mothers are expected to face some level of nipple issues during breastfeeding, which, if not treated, may cause an average of 35% of these mothers to cease breastfeeding before 1 month [28,45]. Nipple damage is painful and may be visible or invisible. When visible, it can present features at the skin surface, such as fissures, cracks, pus, blood, scarring, or crusting. Some skin dermatoses, such as thrush, herpes, eczema, and psoriasis, are also responsible for discomfort and pain during breastfeeding. These conditions can be caused by friction, weather, and temperature changes and using medications or ingredients that can make the skin prone to these disorders. Dermatoses conditions present on both breast and nipple and can have visible features such as scarring, crusting formations, redness, and thickened skin regions [44]. Our research incorporates breast and nipple images from the following disorders: breast abscess, dermatoses, breast engorgement, inflammatory and bacterial mastitis, nipple blebs, and nipple damage.

Current Research Supporting Lactating Mothers

Extensive literature has highlighted the efficacy of deep learning in assessing breast images, helping detect malignant and benign breast tumors for both lactating and nonlactating women [49-54]. This has helped improve the precision of breast ultrasound and mammogram examinations, involving the use of medical imaging previously taken in medical facilities to enhance the evaluation of breast-related illnesses and allow better accuracy in diagnosis for medical personnel [53]. However, these studies relied on images gathered from specialized equipment found only in health care facilities. They did not extend their evaluation to external body images, focusing primarily on aiding health care practitioners in diagnosis. Our work diverges from previous contributions by primarily focusing on using external breast images gathered from personal devices, such as smartphones or cameras from lactating patients, to identify breastfeeding-related conditions in the early stages and evaluate the necessity of further examination and medical intervention.

In the context of breastfeeding disorders, there is a lack of research regarding using deep learning algorithms to evaluate real breast images and identify abnormalities such as mastitis, nipple fissures, dermatoses, and abscesses. To illustrate, literature addressing the early prediction of mastitis mainly originates from agricultural studies, in which the risk of mastitis is constantly assessed to prevent a reduction in animal milk production, which significantly impacts the dairy industry [55,56]. This shows a need for research to adapt these technologies for detecting and preventing breastfeeding disorders in humans. Our study is crucial in settings where access to medical professionals and LCs is limited, as it can help prevent breastfeeding cessation, promote maternal-infant bonding, and improve the overall health and well-being of mothers and infants.

Methods

In this section, we detail the data set collection process, including inclusion and exclusion criteria, data sources, and the characteristics of the images. The section also discusses the artificial intelligence (AI) algorithms used in the study, including the models and their training and validation process, and performance metrics used during evaluation.

Ethical Considerations

This study was approved by the University of California, San Diego Institutional Review Board (801,904). We did not incorporate any personally identifiable data from the participants into this research.

Data Set Collection

Overview

This study used a breast image data set (refer to [Textbox 1](#) and [Table 1](#)), a compilation of physical and digital images specifically curated to train and validate our deep learning model's ability to distinguish between healthy and unhealthy lactating breasts. The data set includes images categorized according to their respective conditions: healthy lactating breast; nipple injuries due to various causes; nipple blebs due to plugged ducts; breast or nipple with signs of dermatoses; and breasts with engorgement, mastitis, or abscess.

Textbox 1. Data set description.**Description**

- Data set size
 - 393.7 MB (each image: minimum 0.015, average 0.360, and maximum 3.575 MB)
- Dimensions (pixels)
 - Width (minimum 68, average 606, and maximum 2448)
 - Height (minimum 68, average 607, and maximum 2448)
- Number of images
 - 1078
- Number of classes
 - 7
- Number of unique subjects
 - 586
- Number of images per class
 - Abscess: 115
 - Dermatoses: 123
 - Engorgement: 63
 - Mastitis: 180
 - Nipple bleb: 82
 - Nipple damage: 197
 - Healthy: 318
- Visual features per class
 - Abscess: swelling and redness, area with palpable fluid collection, and pus
 - Dermatoses: rash, discoloration, flaky skin, uneven skin tone, crusting, and redness
 - Engorgement: swelling, redness, skin stretched and shiny, and enlarged nipple
 - Mastitis: red patches on breast or nipple, swelling, and pus or blood discharge
 - Nipple bleb: small white or yellow bumps on nipple or areola, similar to a blister
 - Nipple damage: nipple swelling, redness, peeling or flaking skin, bleeding, and shape differences
 - Healthy: regular breast and nipple color, may have visible veins
- Number of images per source
 - Physical: 178 (eg, books, magazines, and articles)
 - Physician websites: 366
 - YouTube: 65 (eg, educational channels on women's health)
 - Other: 469 (eg, received by lactation consultants; international board-certified lactation consultant's Instagram, Google Images, and Flickr; support groups mediated by lactation consultants on social media; and other educational websites)

Table 1. Number of images per skin tone per class (FST^a [57]).

Class name	FST I	FST II	FST III	FST IV	FST V	FST VI	Not classified ^b
Abscess	28	35	20	8	14	8	2
Dermatoses	17	37	48	13	3	3	2
Engorgement	4	6	18	30	4	0	1
Mastitis	44	69	51	11	1	4	0
Nipple bleb	9	16	18	8	6	3	22
Nipple damage	40	59	22	15	11	5	45
Healthy	61	90	92	21	28	21	5
Total per FST	203	312	269	106	67	44	77

^aFST: Fitzpatrick skin type.

^bNot classified due to the absence of breast tissue around the nipple in the image.

Data Inclusion and Exclusion Criteria

To be included in the data set, images must meet the following criteria: (1) the image must be in red, green, and blue (RGB) format, either as PNG or JPEG; (2) it must visually have at least 1 of the 7 conditions; (3) the breast or nipple should be visible; (4) the image should be hosted in a trustworthy source (ie, from medical professionals such as physicians, midwife nurses, and IBCLCs), in which the image must have a word or description identifying its condition among the 7 classes to be included as its label; and (5) the visual condition present in the image and the label provided describing the condition should match. Images were excluded from the data set if (1) the breast or nipple were from nonlactating female patients; (2) the condition described on the label and the visual features of the image did not match; (3) the breast or nipple was not visible in the image; and (4) the image did not have any label describing it. A board-certified nurse practitioner (ie, Certified Nurse Practitioner, Advanced Registered Nurse Practitioner, or IBCLC) with >15 years of experience performed a final review of the data set to ensure that images and labels had no discrepancies.

Data Source

We collected images from diverse sources such as breastfeeding-related books, articles, web-based blogs for mothers and physicians, YouTube videos from educative organizations, and social media platforms (eg, Instagram,

Facebook, and Twitter) of certified health care providers who would have educative resources for mothers. To ensure diversity in geographic and racial representation, we conducted image searches using multiple languages (eg, English, Portuguese, Spanish, French, and Chinese) and used search engines adjusted for other countries.

The images were obtained from a diverse group of female patients with several skin colors and breast and nipple sizes, with unstandardized image sizes, orientations, backgrounds, and light sources. In total, the data set consisted of 1078 images, with 318 images of healthy breasts, 115 images of breast abscesses, 123 images of dermatoses, 63 images of breast engorgement, 180 images of mastitis, 82 images of nipple blebs, and 197 images of nipple damage. As shown in Figure 1 and Table 1, a healthy lactating breast presented a uniform color, was free of redness, and had no signs of discharge. Nipples were expected to exhibit a variety of shapes, including flat, protruded, or inverted, and to vary in size. In engorgement, images showed breast and nipple swelling, skin stretched and shiny, and some light redness due to high milk production. For nipple blebs or nipple damage, signs of laceration, blood, blisters, and redness were expected. Mastitis showed swelling, redness, and discharge of pus or blood in the nipple. Abscess shared similarities with mastitis but involved worsened redness and pus in the infected region and may display signs of rupture. Finally, dermatosis images contained signs of skin rash, breast or nipple uneven skin tone, and crusting.

Figure 1. Example images from the testing set that were correctly classified and show features of each breastfeeding-related condition: (A) abscess, (B) dermatoses, (C) engorgement, (D) mastitis, (E) nipple bleb, (F) nipple damage, and (G) healthy.



AI Algorithms

We examined the performance of 5 CNNs commonly used in computer vision problems: Visual Geometry Group model with 16 layers (VGG16) [58], Resnet50 [59], InceptionV3 [60], EfficientNetV2 [61], and DenseNet169 [62]. All models were

built with the PyTorch library for image classification, in which the models had all layers frozen except for the last layer, which was replaced with a fully connected layer adapted to the number of classes—2 for binary classification and 7 for the multiclass task. All models were trained for 100 epochs using the AdamW optimizer with a learning rate of 3e-4, weight decay of 0.1, and

batch size of 20. We chose 100 epochs because it was a converging point where the accuracy no longer increased or decreased. For the loss functions, we applied Binary Cross-Entropy with Logits Loss for binary classification tasks, and for multiclass tasks, we used Cross-Entropy Loss, both fine-tuned with class weights to strategically adjust for class imbalances by proportionally penalizing misclassifications in less represented classes. These models were evaluated using stratified k-fold cross-validation with 10 folds. To ensure the robustness of our cross-validation process, we reset any learned parameters by initializing the models from scratch at the beginning of each fold. Instead of using the entire image data set to train the model, we did feature extraction to optimize the training process (detailed in the Feature Extraction section). We compared the performance of the 5 models across the same data and keep the hyperparameters the same: learning rate, weight decay, batch size, and number of epochs.

Data Set Preprocessing

Before using the images as inputs for the deep learning models, the images were manually cropped to ensure they were deidentified and had no irrelevant content, such as unrelated body areas, clothes, jewelry, identifiable tattoos, or backgrounds, enhancing the model's accuracy and performance. The images were cropped in a 1:1 ratio to prevent image flattening or warping during resizing and loss of important features. Most images have breast and nipple tissue concentrated in the center of the image, thereby focusing the model's evaluation on the most relevant areas. Our image preprocessing guidelines followed similar works in dermatology for AI disease detection and telehealth applications [63-65], which aim to objectively show the area of interest for optimized detection and reduce risks of poorly triaged images.

After cropping the images in a 1:1 ratio and before entering the deep learning pipeline, we applied some standard transformations in the data, starting with image resizing. In this paper, we trained, validated, and tested our data set using 5 different models. Notably, 4 of the chosen models (VGG16, Resnet50, EfficientNetV2, and DenseNet169) specified the input images to be resized to 224×224 pixels, and the InceptionV3 model required input images to be resized to 299×299 pixels. Therefore, we proceeded with the image

resizing according to each model's requirements. The last transformation step incorporates normalization of the images, a procedure where the pixel intensity values are standardized across the data set. To help the models generalize better for our data set, we calculated the mean and SD of all images in the data set to use in the normalization process instead of using the ImageNet data set pretrained parameters, inspired by the previous work involving skin disease classification [66].

Data Set Augmentation

In the process of curating the data set, we recognized that the number of images per class was constrained, given the complexity of gathering images and variability in the clinical features of each class. We implemented data augmentation techniques to mitigate these limitations, reduce the risk of overfitting, and enrich the data set. These techniques artificially expanded the data set by generating realistic transformations of the existing images. We implemented the following 6 data augmentations that were previously used in data sets involving skin lesions [63,67]: center zoom, random rotation, brightness, shear, vertical flip, and horizontal flip. Samples of augmentation are shown in Figure 2. Before data augmentation, our data set consisted of 1078 images. After the augmentation, the data set consisted of 6478 images. The detailed number of samples before and after augmentation is shown in Table 2.

We evaluated our data set before and after data augmentation. In the original data set, the 1000 images were allocated for training and validation, split using stratified k-fold cross-validation [68] with 10 folds. In this process, 90% (900/1000) of the data are used for training and 10% (100/1000) for validation within each fold, as described in Figure 3. The stratified k-fold maintains the proportion of images in each class in both train and validation splits, making sure each fold will be representative of the overall data set. The remaining 78 images were completely excluded from these folds and reserved exclusively for final testing to assess the model's performance on unseen data. After augmenting the original data set, we expanded it to 6000 images for training and validation. Similarly, we increased our test set to 468 images to maintain consistency with the expanded training data, ensuring the model's evaluation on unseen examples remains robust.

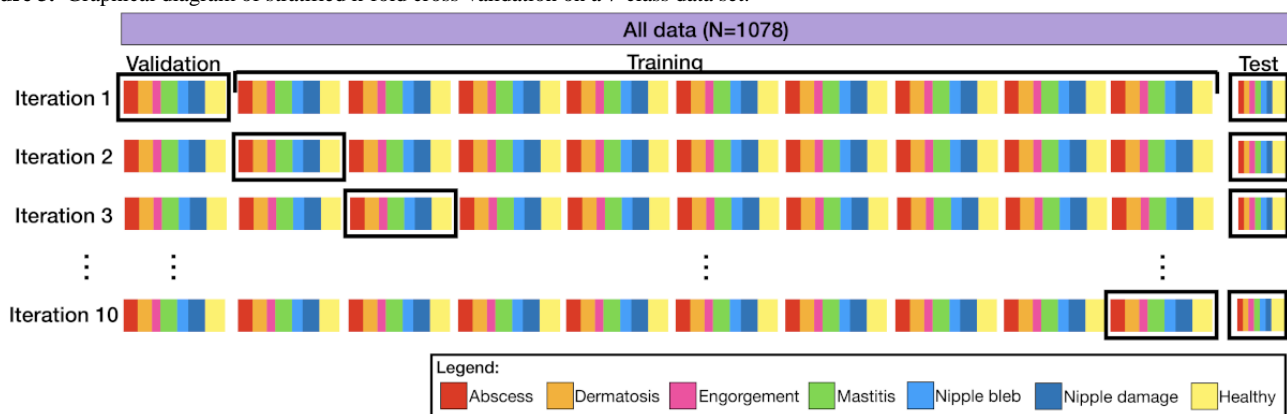
Figure 2. Samples of augmented data: (A) original, (B) brightness, (C) center zoom, (D) horizontal flip, (E) rotation, (F) shear, and (G) vertical flip.



Table 2. Detailed number of samples in the data set.

Data set and classes	Train samples, n	Test samples, n	Train samples (augmented), n	Test samples (augmented), n
7-class data set				
Abscess	108	7	648	42
Dermatoses	115	8	690	48
Engorgement	55	8	330	48
Mastitis	171	9	1026	54
Nipple bleb	75	7	450	42
Nipple damage	188	9	1128	54
Healthy	288	30	1728	180
Binary data set				
Unhealthy ^a	657	40	3942	240
Healthy ^a	343	38	2058	228

^aUnhealthy class combines the classes abscess, dermatoses, mastitis, nipple bleb, and nipple damage, while the healthy class combines healthy and engorgement, all from the 7-class data set.

Figure 3. Graphical diagram of stratified k-fold cross-validation on a 7-class data set.

Feature Extraction

We performed feature extraction using 5 models pretrained on the ImageNet data set. This process helped to reduce the number of computational resources necessary for processing the data set by transforming images into numerical features, without losing relevant information. The models were set to evaluation mode, in which the feature maps are extracted from the final convolutional layers. These maps were then processed through adaptive pooling and flattened into 1D arrays. The extracted features were saved and used as input for the model classifiers.

Training and Evaluation

As previously mentioned in the AI Algorithms section, a total of 5 CNNs were trained on the data set. We proposed 4 tasks in this study, which evaluates the CNNs in the following data sets: (1) multiclass not augmented, (2) multiclass augmented, (3) binary not augmented, and (4) binary augmented. As described in Table 2, we performed an additional 2 evaluations considering a binary model to assess the models' capacity to differentiate between healthy and unhealthy images. The unhealthy class consolidates 5 of the previous conditions: abscess, dermatoses, mastitis, nipple bleb, and nipple damage.

The healthy class consolidates the original healthy and engorgement conditions. For this binary evaluation, we included engorgement images in the healthy condition because it is not inherently indicative of disease and often resolves without medical intervention. Furthermore, engorgement shares visual characteristics with healthy breast conditions, which might not be distinguishable at an early, nonproblematic stage. All models underwent k-fold cross-validation, where we collected performance metrics from each fold and computed their average. We assessed the models' performance for the multiclass and binary data sets using the same metrics: accuracy, precision, recall, F₁-score, and the receiver operating characteristic AUC (ROC-AUC).

Results

Overview

We collected 1078 unique breast images from the web and physical resources, 1000 images as part of the training and validation set, and 78 images as part of the testing set. The augmented data set has 6000 images for training and validation and 468 images for testing. In the *Multiclass Image Detection*

Evaluation section, we show evaluation results from the multiclass and binary data sets, which we evaluated before and after data augmentation. There was no hyperparameter tuning between each fold, and all models had the same optimizer, learning rate, weight decay, and batch size.

Multiclass Image Detection Evaluation

We evaluated 5 CNNs on their ability to distinguish between healthy and 6 breastfeeding-related issues. Table 3 presents the aggregated evaluation metrics for each model sorted based on the test accuracy. The precision, recall, F_1 -score, and overall area under the ROC-AUC are reported as weighted averages to account for the class imbalance within the data sets, ensuring that each class contributes to the final metric in proportion to its prevalence. For each fold in the cross-validation, a separate test set was used to evaluate the model, and the metrics presented are the mean of these evaluations. The best-performing model was Resnet 50, as it managed to contain the best testing accuracy, followed by VGG16 and EfficientNetV2 on a small performance difference. With a similar weighted average setting, in a one-versus-rest fashion, the models achieved an overall ROC-AUC of 0.934 for VGG16, 0.929 for Resnet50, 0.912 for InceptionV3, 0.908 for Densenet169, and 0.872 for EfficientNetV2. The detailed ROC-AUC per class for each model is shown in Figure 4.

When applying data augmentation to the multiclass model, we provided a wider variety of images to help the model better generalize from the training data while not altering the original class distribution. In Figure 5 and Table 4, we show the results across the CNNs after data augmentation, where most of the models showed improved metrics, with Resnet50 being the leading model. The models achieved a ROC-AUC of 0.934 for

Resnet50, 0.912 for VGG16, 0.909 for Densenet169, 0.898 for InceptionV3, and 0.893 for EfficientNetV2.

Looking into the performance of the best model, the Resnet50 with the augmented data set, we can look closer at the metrics per class of this CNN. Table 5 shows the results for 10-fold cross-validation, in which the model had an overall consistent performance across the iterations. Figure 6 presents the aggregated confusion matrix for the Resnet50 model, in which we consolidated the predictions across all 10 iterations applied to the augmented data set. We achieved this aggregation by taking the median predicted class for each instance over the multiple folds, synthesizing a singular prediction representing the consensus of the model's behavior across the test set.

Out of the 468 images used in the testing set, the model could correctly classify 341 images. The total images correctly classified by category are as follows: abscess (24/42; accuracy=57%), dermatoses (43/48; accuracy=90%), engorgement (25/48; accuracy=52%), mastitis (26/54; accuracy=48%), nipple bleb (30/42; accuracy=71%), nipple damage (41/54; accuracy=76%), and healthy (152/180; accuracy=84%). The remaining images that were incorrectly classified happened throughout visually similar conditions and the conditions that can precede each other. Table 6 summarizes the selected model's performance per class on the augmented test set. The model had difficulty categorizing between abscesses, which had false positives on dermatoses and mastitis for 12% (5/42) and 19% (8/42) of the images, respectively. Breast engorgement had false positives on mastitis and healthy breasts for 15% (7/48) and 33% (16/48) of the images, respectively. Mastitis had false positives in abscess (12/54, 22%), nipple damage (9/54, 17%), and healthy breasts (6/54, 11%). About 21% (9/42) of the nipple bleb images were confused as nipple damage.

Table 3. Average evaluation metrics for the trained models on the not augmented data set (sorted based on performance).

Data set and model	Training accuracy	Validation accuracy	Test set metrics			
			Accuracy	Precision	Recall	F_1 -score
7-class data set						
Resnet50	0.907	0.737	<i>0.608</i> ^a	0.675	<i>0.623</i> ^a	<i>0.637</i> ^a
VGG16 ^b	0.818	0.678	0.604	0.674	0.589	0.600
EfficientNetV2	0.779	0.626	0.604	0.658	0.582	0.593
InceptionV3	0.903	0.727	0.574	<i>0.680</i> ^a	0.607	0.622
DenseNet169	<i>0.932</i> ^a	<i>0.771</i> ^a	0.507	0.659	0.596	0.572

^aItalicized items represent the best metric.

^bVGG16: Visual Geometry Group model with 16 layers.

Figure 4. Performance of the 5 convolutional neural networks on the 7-class data set: (A) Resnet50, (B) Visual Geometry Group model with 16 layers (VGG16), (C) EfficientNetV2, (D) InceptionV3, and (E) DenseNet169. AUC: area under the curve.

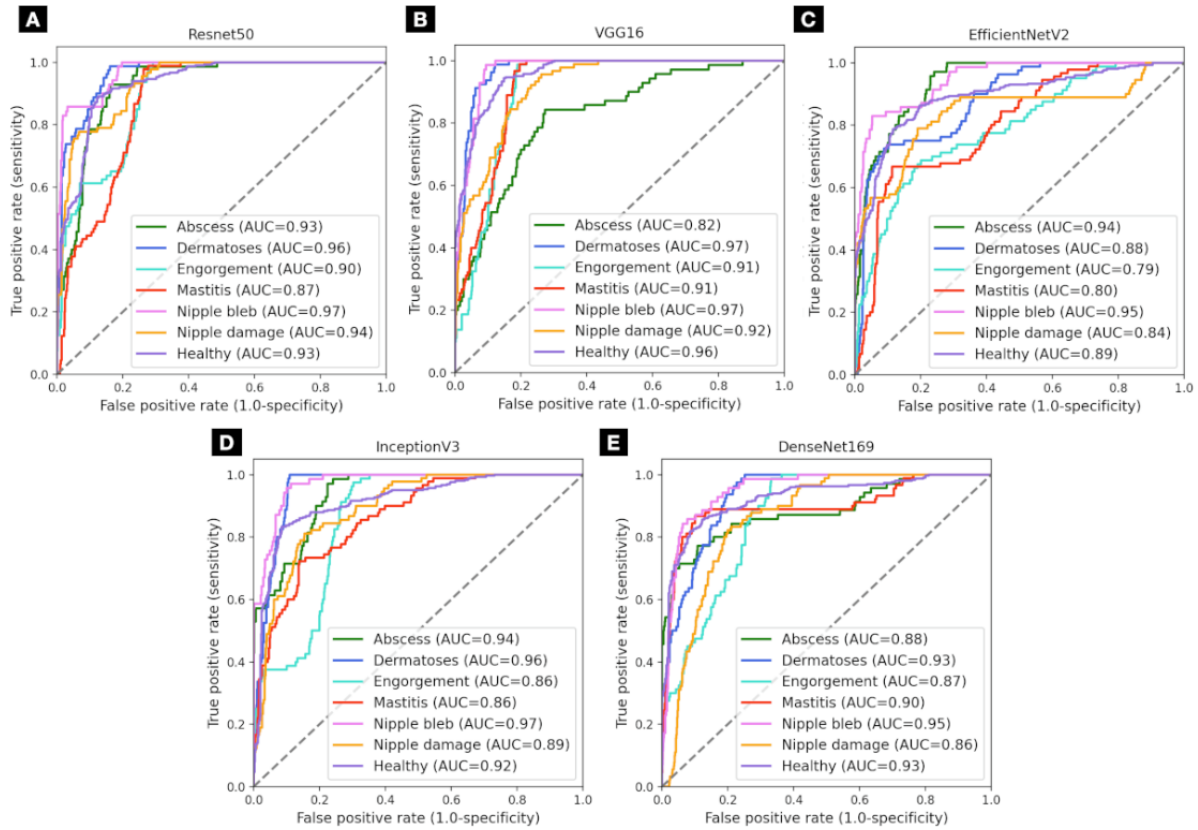


Figure 5. Performance of the 5 convolutional neural networks on the 7-class augmented data set: (A) Resnet50, (B) InceptionV3, (C) EfficientNetV2, (D) Visual Geometry Group model with 16 layers, (E) DenseNet169. AUC: area under the curve.

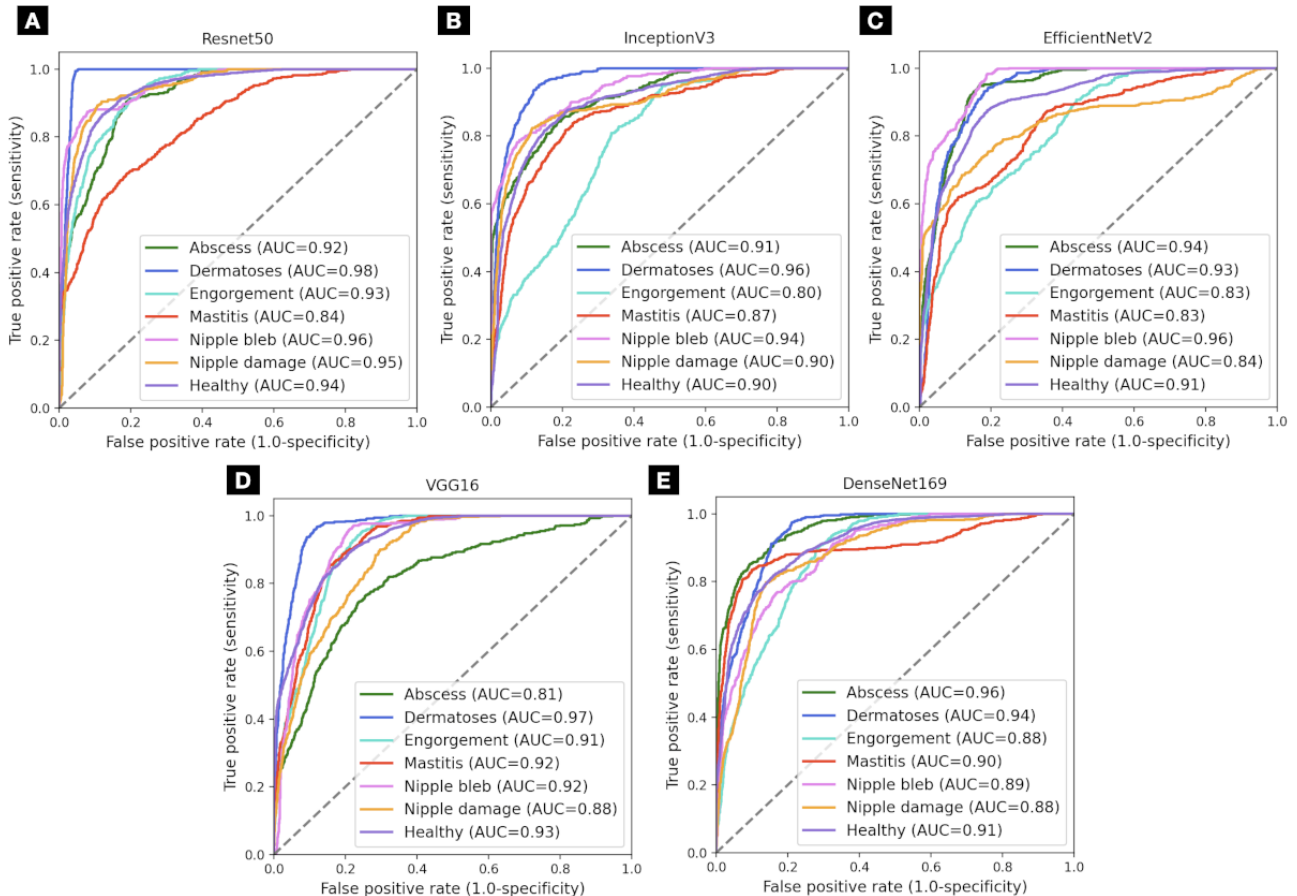


Table 4. Average evaluation metrics for the trained models on the augmented data set (sorted based on performance).

Data set and model	Training accuracy	Validation accuracy	Test set metrics			
			Accuracy	Precision	Recall	F ₁ -score
7-class augmented data set						
Resnet50	0.953	<i>0.907^a</i>	<i>0.672^a</i>	<i>0.717^a</i>	<i>0.715^a</i>	<i>0.713^a</i>
InceptionV3	0.920	0.844	0.617	0.692	0.637	0.649
EfficientNetV2	0.803	0.808	0.602	0.650	0.586	0.5999
VGG16 ^b	0.755	0.801	0.585	0.644	0.561	0.563
DenseNet169	<i>0.954^a</i>	0.889	0.506	0.639	0.611	0.553

^aItalicized items represent the best metric.

^bVGG16: Visual Geometry Group model with 16 layers.

Table 5. Results of 10-fold cross-validation for the augmented data set on Resnet50.

10-fold iterations	Accuracy	Precision	Recall	F ₁ -score
Iteration 1	0.699	0.705	0.699	0.699
Iteration 2	0.714	0.715	0.714	0.712
Iteration 3	0.709	0.713	0.709	0.709
Iteration 4	0.729	0.730	0.729	0.727
Iteration 5	0.718	0.719	0.718	0.716
Iteration 6	0.733	0.734	0.733	0.730
Iteration 7	0.720	0.722	0.720	0.718
Iteration 8	0.707	0.711	0.707	0.706
Iteration 9	0.707	0.707	0.707	0.705
Iteration 10	0.720	0.715	0.720	0.713

Figure 6. Aggregated confusion matrix for the Resnet50 model for the augmented data set with example images from the augmented data set that were correctly and incorrectly classified across all folders.

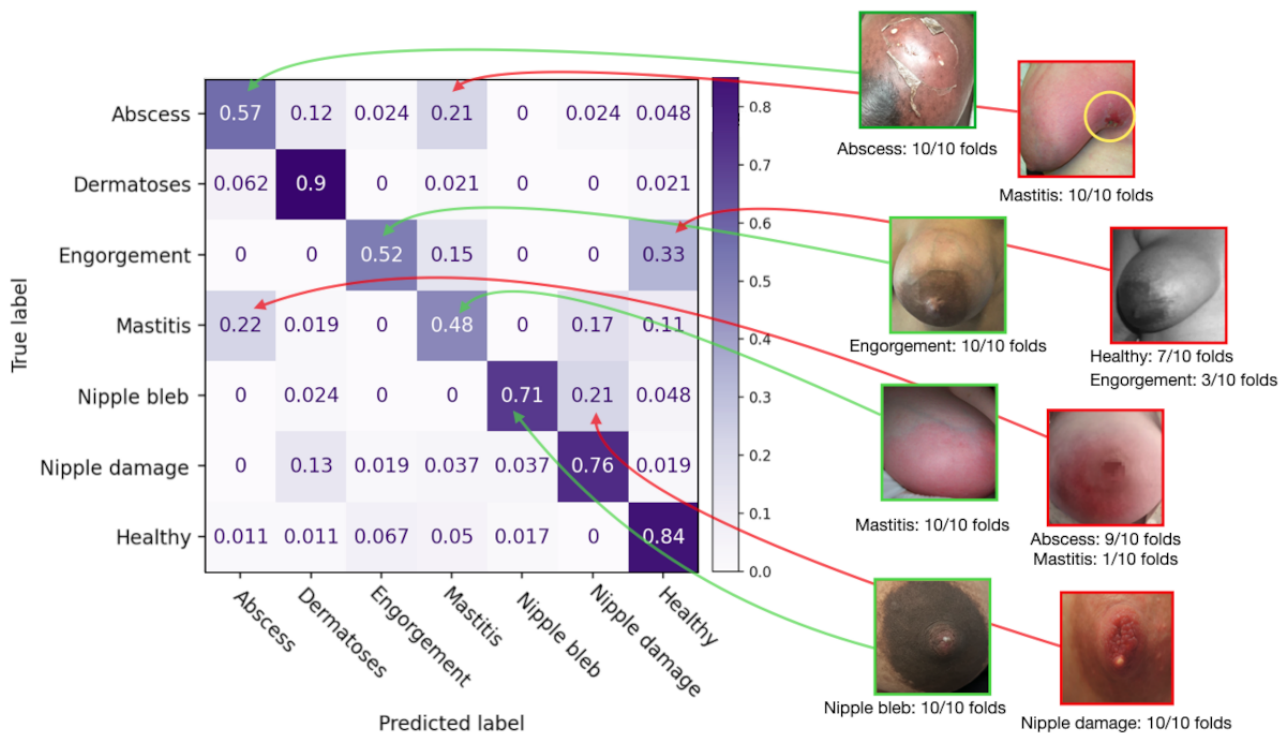


Table 6. Summary of the detection results per class: accuracy, precision, recall, F1-score, and support (ie, number of samples per class) using the Resnet50 architecture.

Class	Accuracy	Precision	Recall	F_1 -score	Support
Abscess	0.571	0.585	0.571	0.578	42
Dermatoses	0.895	0.729	0.895	0.804	48
Engorgement	0.520	0.641	0.520	0.575	48
Mastitis	0.481	0.481	0.481	0.481	54
Nipple bleb	0.714	0.857	0.714	0.779	54
Nipple damage	0.759	0.683	0.759	0.719	54
Healthy	0.844	0.844	0.844	0.844	180

Binary Image Detection Evaluation

To improve the accuracy of our clinical predictions and reduce the chances of incorrect results, we simplified our data set of 7 categories to just 2: healthy and unhealthy. The unhealthy category now includes 5 conditions: abscess, dermatoses, mastitis, nipple bleb, and nipple damage. The healthy category now includes the original healthy conditions and engorgement. Engorgement shares many visual similarities with healthy breast conditions, which made it difficult for the multiclass models to identify engorgement accurately. As presented previously, 33% (16/48) of the images of engorgement were classified as healthy. [Table 7](#) presents the aggregated evaluation metrics for 5 models sorted based on the test accuracy.

The accuracy is reported as a balanced score to address class imbalance, ensuring that each class contributes equally to the final metric. Precision, recall, and F_1 -score are reported for the positive class, with the positive class label specified. For each fold in the cross-validation, we used a separate test set to evaluate the model, and the reported metrics are the average of these evaluations. The best-performing model was the VGG16, which contained the best testing accuracy, followed by Resnet50 and InceptionV3. The models achieved an overall ROC-AUC of 0.977 for VGG16, 0.966 for Resnet50, 0.935 for InceptionV3, 0.921 for EfficientNetV2, and 0.910 for Densenet169. The detailed ROC-AUC for the not augmented and augmented data set is shown in [Figures 7A](#) and [7B](#), respectively.

When applying data augmentation to the binary model, we provided a wider variety of images to help the model better generalize from the training data while not altering the original class distribution. In [Table 8](#), we show the results across the CNNs after data augmentation, where most of the models

showed improved metrics, with Resnet50 being the leading model. The models achieved a ROC-AUC of 0.962 for Resnet50, 0.956 for VGG16, 0.931 for EfficientNetV2, 0.929 for InceptionV3, and 0.915 for Densenet169.

Looking into the performance of the best model, the Resnet50 with the augmented data set, we can look closer at the metrics per class of this CNN. [Table 9](#) shows the results for 10-fold cross-validation, in which the model had an overall consistent performance across the iterations. [Figure 8](#) presents the aggregated confusion matrix for the Resnet50 model, in which we consolidated the predictions across all 10 folds applied to the augmented data set. This aggregation was achieved by taking the median predicted class for each instance over the multiple folds, synthesizing a singular prediction representing the consensus of the model's behavior across the test set.

Out of the 468 images used in the testing set, the model could correctly classify 411 images. The total images correctly classified by category are as follows: unhealthy (228/240; accuracy=95%, precision=83.5%, recall=95% and F_1 -score=89%) and healthy (183/228; accuracy=80.3%, precision=94%, recall=80% and F_1 -score=86.5%). The remaining images that were incorrectly classified presented redness (ie, for engorgement cases misclassified as unhealthy; 26/228), and incomplete images (ie, too close or nipple and breast not fully visible; 12/228). Discussion

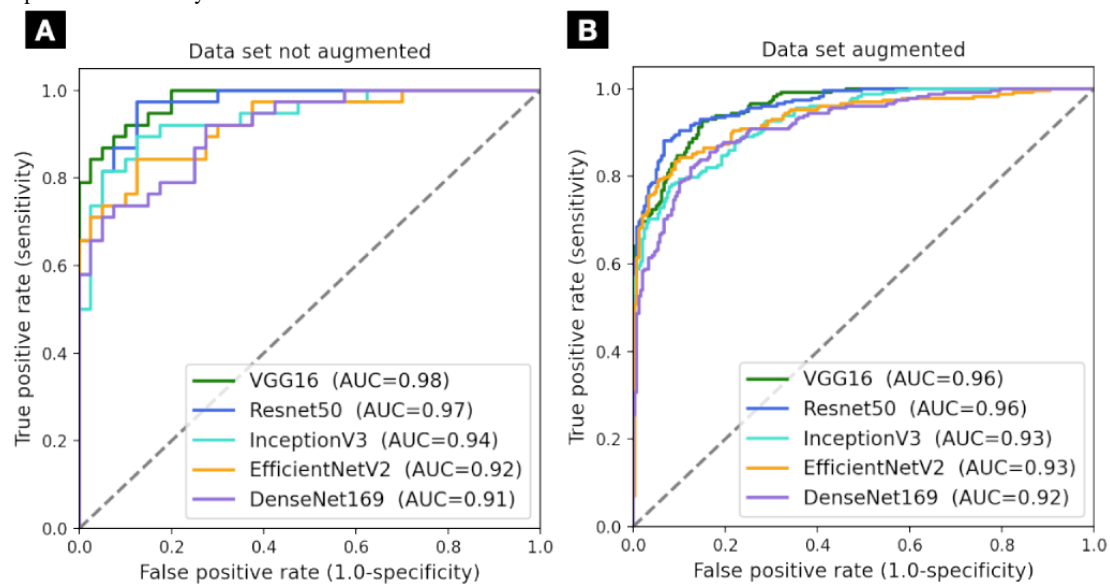
The issues that caused model misclassification included (1) wrong positioning of the breast in the image, (2) common visual features in the images between the classes, (3) a lack of variety of images belonging to specific cases in the data set due to variety limitations, and (4) presence of an extraneous object in the frame. [Figure 1](#) presents the correct prediction from the 7 classes.

Table 7. Average evaluation metrics for the trained models on the not augmented binary data set (sorted based on test accuracy).

Data set and model	Training accuracy	Validation accuracy	Test set metrics			
			Accuracy	Precision	Recall	F_1 -score
Binary data set						
VGG16 ^a	0.901	0.877	<i>0.877</i> ^b	0.990	<i>0.760</i> ^b	<i>0.859</i> ^b
Resnet50	0.923	0.872	0.832	0.954	0.715	0.817
InceptionV3	0.906	0.845	0.838	0.963	0.702	0.812
EfficientNetV2	0.866	0.831	0.811	<i>0.991</i> ^b	0.629	0.769
DenseNet169	<i>0.935</i> ^b	<i>0.880</i> ^b	0.761	0.990	0.529	0.688

^aVGG16: Visual Geometry Group model with 16 layers.

^bItalicized items represent the best metric.

Figure 7. Model performance on the binary data set: (A) without augmentation and (B) with augmentation. AUC: area under the curve; VGG16: Visual Geometry Group model with 16 layers.**Table 8.** Average evaluation metrics for the trained models on the augmented binary data set (sorted based on performance).

Data set and model	Training accuracy	Validation accuracy	Test set metrics			
			Accuracy	Precision	Recall	F_1 -score
Binary augmented data set						
Resnet50	<i>0.952</i> ^a	<i>0.933</i> ^a	<i>0.877</i> ^a	<i>0.941</i> ^a	<i>0.801</i> ^a	<i>0.865</i> ^a
VGG16 ^b	0.877	0.897	0.832	0.941	0.688	0.802
InceptionV3	0.920	0.893	0.831	0.927	0.715	0.807
EfficientNetV2	0.885	0.891	0.825	<i>0.975</i> ^a	0.666	0.791
DenseNet169	0.946	0.927	0.771	0.952	0.570	0.713

^aItalicized items represent the best metric.

^bVGG16: Visual Geometry Group model with 16 layers.

Table 9. Results of 10-fold cross-validation for the augmented binary data set on Resnet50.

Iteration of 10-fold	Accuracy	Precision	Recall	F_1 -score
Iteration 1	0.769	0.948	0.557	0.702
Iteration 2	0.761	0.960	0.531	0.684
Iteration 3	0.791	0.951	0.601	0.737
Iteration 4	0.782	0.970	0.570	0.718
Iteration 5	0.782	0.932	0.596	0.727
Iteration 6	0.778	0.943	0.579	0.717
Iteration 7	0.793	0.928	0.623	0.745
Iteration 8	0.767	0.961	0.544	0.695
Iteration 9	0.778	0.963	0.566	0.713
Iteration 10	0.771	0.969	0.548	0.700

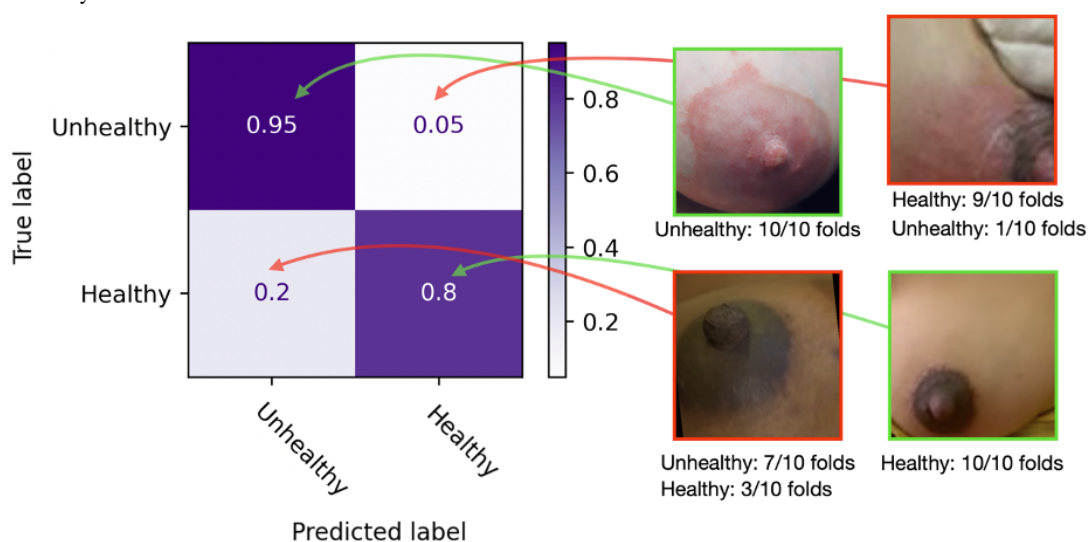
Figure 8. Aggregated confusion matrix for the Resnet50 model for the augmented data set with example images from the augmented data set that were correctly and incorrectly classified across all folders.

Image Quality

When examining misclassification results in our image data set study, we found many image quality issues that likely contributed to the model's diminished performance. In the example images from the testing set, Figures 9A-9C demonstrate good image samples that allow a complete evaluation of the breast's condition and, therefore, can be used for the model's evaluation. These images fully or almost entirely show the nipple at a distance that allows diagnosis and does not show information about the person's surroundings or extraneous objects that the model might misinterpret. In Figures 9D and 9E, the main issue in both examples is the lack of nipple or breast presence or only partial presence, making it difficult for the model to assimilate them with breast figures; even if there are signs of mastitis or engorgement in both images, the image is incomplete. For Figures 9F and 9G, the presence of hands or fingers, nail polish, and partially occluded areas with extraneous objects also affects the model interpretation, especially because we did not train the model with such extra components.

Other issues noted in the preprocessing phase were causing issues in training and validation loss as well as false positive and negative detections. For example, having the image of both breasts instead of one affect prediction accuracy, especially in cases where one breast has a different condition compared to the other. The model did not have a large variety of images showing both breasts. Therefore, we improved the training and test results metrics once we separated the breasts into different figures. In addition, we encountered classification problems with extracted images that show some background components, such as clothes surrounding the breast, breast pumps, or segments of the baby's face or hands. The issues were corrected for these cases by cropping the image to the area of interest. If an object was too similar, such as a hand or a baby, we manually applied blurriness filters in the area and removed saturation so that only the breast is recognizable. Images with low resolution also affect the model's performance, especially if they are originally smaller than the size determined by the data augmentation algorithm and were stretched later. Some images that belonged to this case and were misclassified had their size manually corrected afterward, and the model properly classified them afterward.

Figure 9. Example images from the testing set. (A), (B), and (C) High-quality images, with a full view of the breast and nipple. (D) Image in which the full breast does not appear, making it hard to classify which condition it belongs to. (E) Although the condition is clear and the full breast is visible, the nipple is pixelated in the photo, altering the original features that the model is not used to. (F) and (G) Partially occluded breasts, and the presence of nail polish in the color of the wound also impacts the model's performance in those cases. The examples of low-quality data provide details about how to improve data acquisition for future development.



Visual Similarities Between Conditions

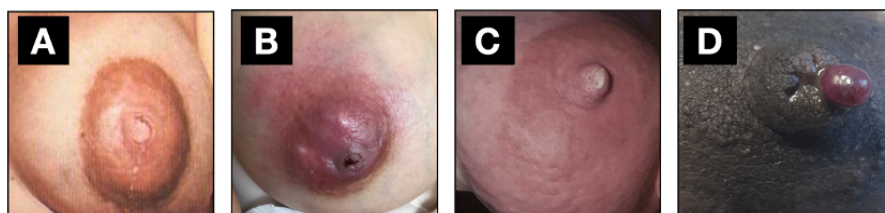
Conditions that present common features and can cause confusion in the diagnosis are mastitis, engorgement, and healthy. Mastitis shows redness throughout the entire breast, showing little skin tone differences and making breasts appear fuller. Some of these features are commonly found in breast engorgement. However, there are fewer signs of intensified redness, sometimes no redness at all, but there may be visible veins and stretched nipples, making them visually similar to healthy ones. Due to the limited availability of images of breast engorgement for a separate class and the fact that engorgement is not necessarily an issue but can become mastitis when not alleviated, the model classified some engorged breasts as mastitis. When we included engorgement in the healthy class for the binary classification, we still got images misclassified as unhealthy, showing how transition conditions should be followed more closely.

This highlights the need for (1) increasing the engorgement data set; (2) working closely with LCs to investigate the need to categorize conditions that can be a problem but indicate false positive cases of more serious issues; and (3) exploring the possibility of using these conditions that have higher errors as a base for following patient condition progression, where there is a transition between conditions for improving or worsening a patient's situation.

Lack of Variety of Images Belonging to Specific Cases in the Data Set

For the case of [Figure 10A](#), the engorged breast occurs in an inverted nipple, showing its center lighter and misclassifying

Figure 10. Images incorrectly classified due to data set variety limitations: (A) an engorged breast with an inverted nipple classified as nipple bleb, (B) breast with an abscess but also has nipple damage, (C) breast with granulomatous mastitis classified as nipple damage, and (D) nipple damage classified as nipple bleb.



Limitations

Our findings emphasize the need for improvement in several areas. As demonstrated in our evaluation, naturalistic images captured by users have several image quality issues that can impede the classification system from proper functioning. Thus,

it as a nipple bleb. Another example of misclassification includes conditions that occur together, which is the case in [Figure 10B](#), showcasing a breast abscess concentrated behind the nipple and with signs of nipple damage. Such an example was one of the very few occurrences of simultaneous conditions in the data set and emphasized the reality that LCs have patients with similar cases, bringing the need to think about systems that (1) recognize multiple conditions or (2) decide between the most severe one for patient priority. [Figure 10C](#) is a case of granulomatous mastitis that was classified as nipple damage due to the presence of nipple scarring, highlighting the fewer occurrences of such a specific case in the data set.

In addition, [Figures 10C](#) and [10D](#) show breasts in the conditions of engorgement and nipple damage, respectively. For [Figure 10D](#), due to the proximity and nature of the nipple damage with a blood blister, the reflection on the dot suggests that it could be a nipple bleb, also misclassifying the image. These misclassified images with distinct features can also be complex to classify for humans, mainly because some of these conditions rarely occur. Given the nature of the images and the lack of images publicly available with the variety of cases across different skin tones, breasts, and nipple sizes, we believe that working with more images involving rare disorders and providing more data augmentation alternatives can improve the model's classification significantly. In addition, [Figure 10D](#) highlights the issue with image angle and proximity. The picture was taken too close to the breast, having a higher chance of misclassification.

future systems must implement a user interface to properly guide parents in taking pictures to input the AI triaging system. This system should provide basic guidelines around how to frame the breast such that no occlusion is present; not use the finger to point out parts of interest; and ensure the camera framing can see the entire breast so that the nipple, areola, and

breast tissue are all visible. Previous works explore the importance of implementing guidelines for image assessment of external diseases, such as in dermatology disease assessments, and its benefits for better professional evaluation and higher accuracy in diagnosing conditions [64,65,69]. Guidelines may be implemented as a set of easy instructions, and more advanced systems could provide immediate image quality feedback.

Moreover, our system only uses RGB images to triage breastfeeding-related conditions, not incorporating patient input regarding pain onset, location, symptoms, and pain levels. These are critical data for diagnosing with higher accuracy and providing more effective feedback to patients experiencing breastfeeding-related pain [70]. Furthermore, automating patient responses [71-73] and using large language models [74] can help categorize issues based on their problem description and image inputs, streamlining the care process and ensuring prompt patient attention.

Finally, the most significant limitation of this work is how this evaluation was limited in having a properly balanced data set to help achieve close-to-perfect performance scores from the model. Despite these limitations, we addressed imbalance issues and proved it possible to obtain satisfactory results in detecting and differentiating the conditions we tested.

Applications and Future Work

This study showcases the potential for high-accuracy breastfeeding-related condition detection to manage postpartum challenges better. In addition, we demonstrate the feasibility of implementing patient support and condition triaging for smartphone-based apps by using deep learning RGB image recognition. The model can be integrated into a telehealth pipeline for postpartum lactation care, helping LCs classify and organize patients based on the severity of their condition or the level of certainty regarding their health concerns. In addition, the system can help track patient disease progression and aid newly qualified LCs by providing faster decision-making support.

The evaluation will serve as a baseline for performing a co-design study with mothers and LCs to evaluate the system requirements regarding data gathering and privacy concerns regarding sensitive data sharing. Understanding the benefits of such a system and recognizing its challenges is essential for building effective tools that will meet patients' and health care providers' needs. Furthermore, a comprehensive approach is

needed to determine the threshold for flagging a patient as unhealthy in the AI-mediated lactation care system, combining quantitative measures (eg, image detection and pain assessment) with clinical expertise. These improvements will allow this work to compose applications for (1) patient self-assessment tools for actionable feedback for breastfeeding pain, (2) reliably identifying cases that require immediate attention and flagging them for LCs, and (3) enabling timely interventions and improved patient outcomes in lactation care. Future work could envision a fully developed hybrid remote consultation system where patients answer questions for the assessment stage, and images are shared between the patient and provider to visualize the severity of the issue before care is provided. Integrating visual information and pain assessment in remote consultations enhances the diagnostic process and enables LCs to deliver tailored care promptly [75] and help overcome burnout from these professionals.

Conclusions

This study demonstrates the feasibility of AI-mediated detection of breast conditions for lactating women. We took the first step in this domain by using RGB breast images to triage healthy from unhealthy breasts in mastitis spectrum disease conditions such as nipple blebs, engorgement, abscess, and mastitis; nipple damage caused by poor breastfeeding techniques, breast pumps, and other conditions; and dermatoses caused by a variety of conditions. We implemented 5 distinct CNN models to classify images from 2 different data sets, identifying 7 breast conditions and distinguishing between healthy and unhealthy conditions. The evaluation of the models based on our data set demonstrated the feasibility of using CNNs to classify and intervene with patients who seek remote guidance and management of their symptoms. Although this model's performance was good, it can be improved by increasing the variety of images and conditions in the data set and implementing the best practices for image posing for proper image classification, leaving significant room for improvement. The feasibility of this work is the initial step toward building tele-lactation services with better data for LCs. We hope our work will inspire future exploration to apply technologies to help lactation support research that can reach more people globally and investigate ideas beyond laboratory settings. This will allow a more comprehensive understanding of breast health for postpartum mothers and empower them to take proactive steps in maintaining their well-being.

Acknowledgments

The authors thank the Google Health Equity Research Initiative that supported this research through their program to advance health equity research and improve health outcomes for groups disproportionately impacted by health disparities.

Data Availability

The data sets generated and analyzed during this study are not publicly available due to confidentiality reasons but are available from the corresponding author on reasonable request.

Authors' Contributions

JDS conceptualized the research question, acquired the data, analyzed the data, wrote the manuscript, and takes responsibility for the integrity of the data and the accuracy of the data analyses. JME provided guidance and assisted with the cross-validation and data augmentation strategies. KC provided guidance during the study design and material support and data consistency. EJW and VKV provided guidance, data analysis, and technical support during the study. All authors contributed to drafting the paper and its critical revision for important intellectual content.

Conflicts of Interest

None declared.

References

1. Kramer MS, Kakuma R. Optimal duration of exclusive breastfeeding. *Cochrane Database Syst Rev* 2012 Aug 15;2012(8):CD003517 [FREE Full text] [doi: [10.1002/14651858.CD003517.pub2](https://doi.org/10.1002/14651858.CD003517.pub2)] [Medline: [22895934](https://pubmed.ncbi.nlm.nih.gov/22895934/)]
2. Duijts L, Ramadhani MK, Moll HA. Breastfeeding protects against infectious diseases during infancy in industrialized countries. A systematic review. *Matern Child Nutr* 2009 Jul;5(3):199-210 [FREE Full text] [doi: [10.1111/j.1740-8709.2008.00176.x](https://doi.org/10.1111/j.1740-8709.2008.00176.x)] [Medline: [19531047](https://pubmed.ncbi.nlm.nih.gov/19531047/)]
3. Kramer MS, Guo T, Platt RW, Sevkovskaya Z, Dzikovich I, Collet JP, et al. Infant growth and health outcomes associated with 3 compared with 6 mo of exclusive breastfeeding. *Am J Clin Nutr* 2003 Aug;78(2):291-295. [doi: [10.1093/ajcn/78.2.291](https://doi.org/10.1093/ajcn/78.2.291)] [Medline: [12885711](https://pubmed.ncbi.nlm.nih.gov/12885711/)]
4. Kent G. Child feeding and human rights. *Int Breastfeed J* 2006 Dec 18;1:27 [FREE Full text] [doi: [10.1186/1746-4358-1-27](https://doi.org/10.1186/1746-4358-1-27)] [Medline: [17176464](https://pubmed.ncbi.nlm.nih.gov/17176464/)]
5. Dinour LM. Speaking out on "breastfeeding" terminology: recommendations for gender-inclusive language in research and reporting. *Breastfeed Med* 2019 Oct 01;14(8):523-532 [FREE Full text] [doi: [10.1089/bfm.2019.0110](https://doi.org/10.1089/bfm.2019.0110)] [Medline: [31364867](https://pubmed.ncbi.nlm.nih.gov/31364867/)]
6. Breastfeeding report card. Centers for Disease Control and Prevention. 2022. URL: <https://www.cdc.gov/breastfeeding/data/reportcard.htm> [accessed 2023-11-13]
7. Breastfeeding. United Nations International Children's Emergency Fund. URL: <https://data.unicef.org/topic/nutrition/breastfeeding/> [accessed 2023-11-13]
8. Coca KP, Marcacine KO, Gamba MA, Corrêa L, Aranha AC, Abrão AC. Efficacy of low-level laser therapy in relieving nipple pain in breastfeeding women: a triple-blind, randomized, controlled trial. *Pain Manag Nurs* 2016 Aug;17(4):281-289 [FREE Full text] [doi: [10.1016/j.pmn.2016.05.003](https://doi.org/10.1016/j.pmn.2016.05.003)] [Medline: [27363734](https://pubmed.ncbi.nlm.nih.gov/27363734/)]
9. Brown CR, Dodds L, Legge A, Bryanton J, Semenic S. Factors influencing the reasons why mothers stop breastfeeding. *Can J Public Health* 2014 May 09;105(3):e179-e185 [FREE Full text] [doi: [10.17269/cjph.105.4244](https://doi.org/10.17269/cjph.105.4244)] [Medline: [25165836](https://pubmed.ncbi.nlm.nih.gov/25165836/)]
10. Friesen CA, Hormuth LJ, Petersen D, Babbitt T. Using videoconferencing technology to provide breastfeeding support to low-income women: connecting hospital-based lactation consultants with clients receiving care at a community health center. *J Hum Lact* 2015 Nov;31(4):595-599. [doi: [10.1177/0890334415601088](https://doi.org/10.1177/0890334415601088)] [Medline: [26297347](https://pubmed.ncbi.nlm.nih.gov/26297347/)]
11. Chaves AF, Vitoriano LN, Borges FL, Alves Melo RD, de Oliveira MG, Chagas Costa Lima AC. Percepção das mulheres que receberam consultoria em amamentação. *Enfermagem em Foco* 2019;10(5). [doi: [10.21675/2357-707X.2019.v10.n5.2519](https://doi.org/10.21675/2357-707X.2019.v10.n5.2519)]
12. Patel S, Patel S. The effectiveness of lactation consultants and lactation counselors on breastfeeding outcomes. *J Hum Lact* 2016 Aug;32(3):530-541. [doi: [10.1177/0890334415618668](https://doi.org/10.1177/0890334415618668)] [Medline: [26644419](https://pubmed.ncbi.nlm.nih.gov/26644419/)]
13. Current statistics on worldwide IBCLCs. International Board of Lactation Consultant Examiners. URL: <https://ibclce.org/about-ibclce/current-statistics-on-worldwide-ibclcs/> [accessed 2023-11-13]
14. Hamilton BE, Martin JA, Osterman MJ. Births: provisional data for 2021. Centers for Disease Control and Prevention. 2022 May. URL: <https://www.cdc.gov/nchs/data/vsrr/vsrr020.pdf> [accessed 2024-06-02]
15. Registros. Portal da Transparência. URL: <https://transparencia.registrocivil.org.br/registros> [accessed 2023-11-13]
16. DeLeo A, Geraghty S. iMidwife: midwifery students' use of smartphone technology as a mediated educational tool in clinical environments. *Contemp Nurse* 2018 Dec 18;54(4-5):522-531 [FREE Full text] [doi: [10.1080/10376178.2017.1416305](https://doi.org/10.1080/10376178.2017.1416305)] [Medline: [29228874](https://pubmed.ncbi.nlm.nih.gov/29228874/)]
17. Tripp N, Hainey K, Liu A, Poulton A, Peek M, Kim J, et al. An emerging model of maternity care: smartphone, midwife, doctor? *Women Birth* 2014 Mar;27(1):64-67. [doi: [10.1016/j.wombi.2013.11.001](https://doi.org/10.1016/j.wombi.2013.11.001)] [Medline: [24295598](https://pubmed.ncbi.nlm.nih.gov/24295598/)]
18. Feinstein J, Slora EJ, Bernstein HH. Telehealth can promote breastfeeding during the COVID-19 pandemic. *NEJM Catal Innov Care Deliv* 2021;2(2):1-11. [doi: [10.1056/CAT.21.0076](https://doi.org/10.1056/CAT.21.0076)]
19. Haase B, Brennan E, Wagner CL. Effectiveness of the IBCLC: have we made an impact on the care of breastfeeding families over the past decade? *J Hum Lact* 2019 Aug 17;35(3):441-452 [FREE Full text] [doi: [10.1177/0890334419851805](https://doi.org/10.1177/0890334419851805)] [Medline: [31206324](https://pubmed.ncbi.nlm.nih.gov/31206324/)]
20. Ray KN, Demirci JR, Uscher-Pines L, Bogen DL. Geographic access to international board-certified lactation consultants in Pennsylvania. *J Hum Lact* 2019 Feb 03;35(1):90-99 [FREE Full text] [doi: [10.1177/0890334418768458](https://doi.org/10.1177/0890334418768458)] [Medline: [29969344](https://pubmed.ncbi.nlm.nih.gov/29969344/)]

21. Hoddinott P, Britten J, Pill R. Why do interventions work in some places and not others: a breastfeeding support group trial. *Soc Sci Med* 2010 Mar;70(5):769-778 [FREE Full text] [doi: [10.1016/j.socscimed.2009.10.067](https://doi.org/10.1016/j.socscimed.2009.10.067)] [Medline: [20005617](https://pubmed.ncbi.nlm.nih.gov/20005617/)]
22. de Souza J, Calsinski C, Chamberlain K, Cibrian F, Wang EJ. Investigating interactive methods in remote chestfeeding support for lactation consulting professionals in Brazil. *Frontiers in Digital Health* 2023 Apr 02;5:1-16 [FREE Full text] [doi: [10.3389/fgdh.2023.1143528](https://doi.org/10.3389/fgdh.2023.1143528)] [Medline: [37077406](https://pubmed.ncbi.nlm.nih.gov/37077406/)]
23. Donovan H, Welch A, Williamson M. Reported levels of exhaustion by the graduate nurse midwife and their perceived potential for unsafe practice: a phenomenological study of Australian double degree nurse midwives. *Workplace Health Saf* 2021 Feb;69(2):73-80. [doi: [10.1177/2165079920938000](https://doi.org/10.1177/2165079920938000)] [Medline: [32812841](https://pubmed.ncbi.nlm.nih.gov/32812841/)]
24. Fraser HS, Blaya J. Implementing medical information systems in developing countries, what works and what doesn't. *AMIA Annu Symp Proc* 2010 Nov 13;2010:232-236 [FREE Full text] [Medline: [21346975](https://pubmed.ncbi.nlm.nih.gov/21346975/)]
25. Busch DW, Logan K, Wilkinson A. Clinical practice breastfeeding recommendations for primary care: applying a tri-core breastfeeding conceptual model. *J Pediatr Health Care* 2014;28(6):486-496. [doi: [10.1016/j.pedhc.2014.02.007](https://doi.org/10.1016/j.pedhc.2014.02.007)] [Medline: [24786581](https://pubmed.ncbi.nlm.nih.gov/24786581/)]
26. Kern-Goldberger AR, Srinivas SK. Obstetrical telehealth and virtual care practices during the COVID-19 pandemic. *Clin Obstet Gynecol* 2022 Mar 01;65(1):148-160 [FREE Full text] [doi: [10.1097/GRF.0000000000000671](https://doi.org/10.1097/GRF.0000000000000671)] [Medline: [35045037](https://pubmed.ncbi.nlm.nih.gov/35045037/)]
27. Burns E, Fenwick J, Sheehan A, Schmied V. Mining for liquid gold: midwifery language and practices associated with early breastfeeding support. *Matern Child Nutr* 2013 Jan 09;9(1):57-73 [FREE Full text] [doi: [10.1111/j.1740-8709.2011.00397.x](https://doi.org/10.1111/j.1740-8709.2011.00397.x)] [Medline: [22405753](https://pubmed.ncbi.nlm.nih.gov/22405753/)]
28. Niazi A, Rahimi VB, Soheili-Far S, Askari N, Rahmanian-Devin P, Sanei-Far Z, et al. A systematic review on prevention and treatment of nipple pain and fissure: are they curable? *J Pharmacopuncture* 2018 Sep 30;21(3):139-150 [FREE Full text] [doi: [10.3831/kpi.2018.21.017](https://doi.org/10.3831/kpi.2018.21.017)]
29. Mitoulas LR, Davanzo R. Breast pumps and mastitis in breastfeeding women: clarifying the relationship. *Front Pediatr* 2022;10:856353 [FREE Full text] [doi: [10.3389/fped.2022.856353](https://doi.org/10.3389/fped.2022.856353)] [Medline: [35757121](https://pubmed.ncbi.nlm.nih.gov/35757121/)]
30. Gomez-Pomar E, Blubaugh R. The Baby Friendly Hospital Initiative and the ten steps for successful breastfeeding. A critical review of the literature. *J Perinatol* 2018 Jun 7;38(6):623-632 [FREE Full text] [doi: [10.1038/s41372-018-0068-0](https://doi.org/10.1038/s41372-018-0068-0)] [Medline: [29416115](https://pubmed.ncbi.nlm.nih.gov/29416115/)]
31. VanDevanter N, Gennaro S, Budin W, Calalang-Javiera H, Nguyen M. Evaluating implementation of a baby friendly hospital initiative. *MCN Am J Matern Child Nurs* 2014;39(4):231-237. [doi: [10.1097/NMC.000000000000046](https://doi.org/10.1097/NMC.000000000000046)] [Medline: [24978002](https://pubmed.ncbi.nlm.nih.gov/24978002/)]
32. Most popular messaging apps worldwide 2023. Similarweb. URL: <https://www.similarweb.com/blog/research/market-research/worldwide-messaging-apps/> [accessed 2023-06-22]
33. Coelho LS. Telefarmácia na atenção primária à saúde: relato de experiência sobre a implementação e prática em um centro de Saúde de Florianópolis. Universidade Federal de Santa Catarina. 2021 Sep 23. URL: <https://repositorio.ufsc.br/handle/123456789/228419> [accessed 2024-06-02]
34. Weaver NS, Roy A, Martinez S, Gomanie NN, Mehta K. How WhatsApp is transforming healthcare services and empowering health workers in low-and middle-income countries. In: *Proceedings of the IEEE Global Humanitarian Technology Conference (GHTC)*. 2022 Presented at: GHTC 2022; September 8-11, 2022; Santa Clara, CA. [doi: [10.1109/ghtc55712.2022.9911048](https://doi.org/10.1109/ghtc55712.2022.9911048)]
35. Trude AC, Martins RC, Martins-Silva T, Blumenberg C, Carpena MX, Del-Ponte B, et al. A WhatsApp-based intervention to improve maternal social support and maternal-child health in southern Brazil: the text-message intervention to enhance social support (TIES) feasibility study. *Inquiry* 2021 Oct 08;58:469580211048701 [FREE Full text] [doi: [10.1177/00469580211048701](https://doi.org/10.1177/00469580211048701)] [Medline: [34619999](https://pubmed.ncbi.nlm.nih.gov/34619999/)]
36. de Araujo JC, de Sousa Lima T, dos Santos JA, dos Santos Costa E. Use of WhatsApp app as a tool to education and health promotion of pregnant women during prenatal care. *Anais do I Congresso Norte Nordeste de Tecnologias em Saúde*. 2018. URL: <https://revistas.ufpi.br/index.php/connts/article/view/7954/4682> [accessed 2024-06-02]
37. Lima AC, Chaves AF, Oliveira MG, Lima SA, Machado MM, Oriá MO. Consultoria em amamentação durante a pandemia COVID-19: relato de experiência. *Esc Anna Nery* 2020;24(spe):e20200350. [doi: [10.1590/2177-9465-ean-2020-0350](https://doi.org/10.1590/2177-9465-ean-2020-0350)]
38. Gavine A, Marshall J, Buchanan P, Cameron J, Leger A, Ross S, et al. Remote provision of breastfeeding support and education: systematic review and meta-analysis. *Matern Child Nutr* 2022 Apr;18(2):e13296 [FREE Full text] [doi: [10.1111/mcn.13296](https://doi.org/10.1111/mcn.13296)] [Medline: [34964542](https://pubmed.ncbi.nlm.nih.gov/34964542/)]
39. Nóbrega V, Melo R, Diniz A, Vilar R. As redes sociais de apoio para o Aleitamento Materno: uma pesquisa-ação. *Saúde Debate* 2019;43(121):429-440 [FREE Full text] [doi: [10.1590/0103-1104201912111](https://doi.org/10.1590/0103-1104201912111)]
40. Hinman R, Lawford B, Bennell K. Harnessing technology to deliver care by physical therapists for people with persistent joint pain: telephone and video - conferencing service models. *J Appl Biobehav Res* 2018 Oct 30;24(2):e12150 [FREE Full text] [doi: [10.1111/jabr.12150](https://doi.org/10.1111/jabr.12150)]
41. Candido NL, Marcolino AM, Santana JM, Silva JR, Silva ML. Remote physical therapy during COVID-19 pandemic: guidelines in the Brazilian context. *Fisioterapia em Movimento* 2022 Mar;35(4):e35202 [FREE Full text] [doi: [10.1590/fm.2022.35202](https://doi.org/10.1590/fm.2022.35202)]

42. Giglia R, Cox K, Zhao Y, Binns CW. Exclusive breastfeeding increased by an internet intervention. *Breastfeed Med* 2015;10(1):20-25. [doi: [10.1089/bfm.2014.0093](https://doi.org/10.1089/bfm.2014.0093)] [Medline: [25358119](https://pubmed.ncbi.nlm.nih.gov/25358119/)]
43. Krishnamurti T, Simhan HN, Borrero S. Competing demands in postpartum care: a national survey of U.S. providers' priorities and practice. *BMC Health Serv Res* 2020 Apr 06;20(1):284 [FREE Full text] [doi: [10.1186/s12913-020-05144-2](https://doi.org/10.1186/s12913-020-05144-2)] [Medline: [32252757](https://pubmed.ncbi.nlm.nih.gov/32252757/)]
44. Berens P, Eglash A, Malloy M, Steube AM. ABM clinical protocol #26: persistent pain with breastfeeding. *Breastfeed Med* 2016 Mar;11(2):46-53. [doi: [10.1089/bfm.2016.29002.pjb](https://doi.org/10.1089/bfm.2016.29002.pjb)] [Medline: [26881962](https://pubmed.ncbi.nlm.nih.gov/26881962/)]
45. Douglas P. Re-thinking lactation-related nipple pain and damage. *Womens Health (Lond)* 2022;18:17455057221087865 [FREE Full text] [doi: [10.1177/17455057221087865](https://doi.org/10.1177/17455057221087865)] [Medline: [35343816](https://pubmed.ncbi.nlm.nih.gov/35343816/)]
46. Mitchell KB, Johnson HM, Rodríguez JM, Eglash A, Scherzinger C, Zakarija-Grkovic I, et al. Academy of breastfeeding medicine clinical protocol #36: the mastitis spectrum, revised 2022. *Breastfeed Med* 2022 May;17(5):360-376. [doi: [10.1089/bfm.2022.29207.kbm](https://doi.org/10.1089/bfm.2022.29207.kbm)] [Medline: [35576513](https://pubmed.ncbi.nlm.nih.gov/35576513/)]
47. Pevzner M, Dahan A. Mastitis while breastfeeding: prevention, the importance of proper treatment, and potential complications. *J Clin Med* 2020 Jul 22;9(8):2328 [FREE Full text] [doi: [10.3390/jcm9082328](https://doi.org/10.3390/jcm9082328)] [Medline: [32707832](https://pubmed.ncbi.nlm.nih.gov/32707832/)]
48. Nakamura M, Asaka Y, Ogawara T, Yorozu Y. Nipple skin trauma in breastfeeding women during postpartum week one. *Breastfeed Med* 2018 Sep;13(7):479-484 [FREE Full text] [doi: [10.1089/bfm.2017.0217](https://doi.org/10.1089/bfm.2017.0217)] [Medline: [30074830](https://pubmed.ncbi.nlm.nih.gov/30074830/)]
49. Aldhyani TH, Nair R, Alzain E, Alkahtani H, Koundal D. Deep learning model for the detection of real time breast cancer images using improved dilation-based method. *Diagnostics (Basel)* 2022 Oct 16;12(10):2505 [FREE Full text] [doi: [10.3390/diagnostics12102505](https://doi.org/10.3390/diagnostics12102505)] [Medline: [36292194](https://pubmed.ncbi.nlm.nih.gov/36292194/)]
50. Yoon JH, Kim EK. Deep learning-based artificial intelligence for mammography. *Korean J Radiol* 2021 Aug;22(8):1225-1239 [FREE Full text] [doi: [10.3348/kjr.2020.1210](https://doi.org/10.3348/kjr.2020.1210)] [Medline: [33987993](https://pubmed.ncbi.nlm.nih.gov/33987993/)]
51. Kim SY, Choi Y, Kim EK, Han BK, Yoon JH, Choi JS, et al. Deep learning-based computer-aided diagnosis in screening breast ultrasound to reduce false-positive diagnoses. *Sci Rep* 2021 Jan 11;11(1):395 [FREE Full text] [doi: [10.1038/s41598-020-79880-0](https://doi.org/10.1038/s41598-020-79880-0)] [Medline: [33432076](https://pubmed.ncbi.nlm.nih.gov/33432076/)]
52. Calisto FM, Nunes N, Nascimento JC. BreastScreening: on the use of multi-modality in medical imaging diagnosis. In: *Proceedings of the International Conference on Advanced Visual Interfaces*. 2020 Presented at: AVI '20; September 28-October 2, 2020; Salerno, Italy. [doi: [10.1145/3399715.3399744](https://doi.org/10.1145/3399715.3399744)]
53. Zhou Y, Feng BJ, Yue WW, Liu Y, Xu ZF, Xing W, et al. Differentiating non-lactating mastitis and malignant breast tumors by deep-learning based AI automatic classification system: a preliminary study. *Front Oncol* 2022 Sep 15;12:997306 [FREE Full text] [doi: [10.3389/fonc.2022.997306](https://doi.org/10.3389/fonc.2022.997306)] [Medline: [36185190](https://pubmed.ncbi.nlm.nih.gov/36185190/)]
54. Shen Y, Shamout FE, Oliver JR, Witowski J, Kannan K, Park J, et al. Artificial intelligence system reduces false-positive findings in the interpretation of breast ultrasound exams. *Nat Commun* 2021 Sep 24;12(1):5645 [FREE Full text] [doi: [10.1038/s41467-021-26023-2](https://doi.org/10.1038/s41467-021-26023-2)] [Medline: [34561440](https://pubmed.ncbi.nlm.nih.gov/34561440/)]
55. Abdul Ghafoor N, Sitkowska B. MasPA: a machine learning application to predict risk of mastitis in cattle from AMS sensor data. *AgriEngineering* 2021 Aug 04;3(3):575-584. [doi: [10.3390/agriengineering3030037](https://doi.org/10.3390/agriengineering3030037)]
56. Fadul-Pacheco L, Delgado H, Cabrera VE. Exploring machine learning algorithms for early prediction of clinical mastitis. *Int Dairy J* 2021 Aug;119:105051. [doi: [10.1016/j.idairyj.2021.105051](https://doi.org/10.1016/j.idairyj.2021.105051)]
57. Fitzpatrick TB. The validity and practicality of sun-reactive skin types I through VI. *Arch Dermatol* 1988 Jun;124(6):869-871. [doi: [10.1001/archderm.124.6.869](https://doi.org/10.1001/archderm.124.6.869)] [Medline: [3377516](https://pubmed.ncbi.nlm.nih.gov/3377516/)]
58. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. *arXiv Preprint posted online September 4, 2014* [FREE Full text] [doi: [10.48550/arXiv.1409.1556](https://doi.org/10.48550/arXiv.1409.1556)]
59. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016 Presented at: CVPR 2016; June 27-30, 2016; Las Vegas, NV. [doi: [10.1109/cvpr.2016.90](https://doi.org/10.1109/cvpr.2016.90)]
60. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. Rethinking the inception architecture for computer vision. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016 Presented at: CVPR 2016; June 27-30, 2016; Las Vegas, NV. [doi: [10.1109/cvpr.2016.308](https://doi.org/10.1109/cvpr.2016.308)]
61. Tan M, Le QV. Efficientnetv2: smaller models and faster training. *arXiv Preprint posted online April 1, 2021* [FREE Full text]
62. Huang G, Liu Z, van der Maaten L, Weinberger KQ. Densely connected convolutional networks. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017 Presented at: CVPR 2017; July 21-26, 2017; Honolulu, HI. [doi: [10.1109/cvpr.2017.243](https://doi.org/10.1109/cvpr.2017.243)]
63. Rafay A, Hussain W. EfficientSkinDis: an EfficientNet-based classification model for a large manually curated dataset of 31 skin diseases. *Biomed Signal Process Control* 2023 Aug;85:104869 [FREE Full text] [doi: [10.1016/j.bspc.2023.104869](https://doi.org/10.1016/j.bspc.2023.104869)]
64. Vodrahalli K, Daneshjou R, Novoa RA, Chiou A, Ko JM, Zou J. TrueImage: a machine learning algorithm to improve the quality of telehealth photos. *Pac Symp Biocomput* 2021;26:220-231 [FREE Full text] [Medline: [33691019](https://pubmed.ncbi.nlm.nih.gov/33691019/)]
65. Finnane A, Curiel-Lewandrowski C, Wimberley G, Caffery L, Katragadda C, Halpern A, et al. Proposed technical guidelines for the acquisition of clinical images of skin-related conditions. *JAMA Dermatol* 2017 May 01;153(5):453-457. [doi: [10.1001/jamadermatol.2016.6214](https://doi.org/10.1001/jamadermatol.2016.6214)] [Medline: [28241182](https://pubmed.ncbi.nlm.nih.gov/28241182/)]

66. Jain S, Singhanian U, Tripathy B, Nasr EA, Aboudaif MK, Kamrani AK. Deep learning-based transfer learning for classification of skin cancer. *Sensors (Basel)* 2021 Dec 06;21(23):8142 [FREE Full text] [doi: [10.3390/s21238142](https://doi.org/10.3390/s21238142)] [Medline: [34884146](https://pubmed.ncbi.nlm.nih.gov/34884146/)]
67. Perez F, Vasconcelos C, Avila S, Valle E. Data augmentation for skin lesion analysis. In: *Proceedings of the Third International Skin Imaging Collaboration Workshop*. 2018 Presented at: ISIC 2018; September 16 and 20, 2018; Granada, Spain. [doi: [10.1007/978-3-030-01201-4_33](https://doi.org/10.1007/978-3-030-01201-4_33)]
68. Kohavi R. A study of cross-validation and bootstrap for accuracy estimation and model selection. In: *Proceedings of the 14th International Joint Conference on Artificial intelligence - Volume 2*. 1995 Presented at: IJCAI'95; August 20-25, 1995; Montreal, QC. [doi: [10.5555/1643031.1643047](https://doi.org/10.5555/1643031.1643047)]
69. Ukharov AO, Shlivko IL, Klemenova IA, Garanina OE, Uskova KE, Mironycheva AM, et al. Skin cancer risk self-assessment using AI as a mass screening tool. *Inform Med Unlocked* 2023;38:101223 [FREE Full text] [doi: [10.1016/j.imu.2023.101223](https://doi.org/10.1016/j.imu.2023.101223)]
70. Lucas R, McGrath J. Clinical assessment and management of breastfeeding pain. *Topics Pain Manag* 2016 Oct;32(3):1-11. [doi: [10.1097/01.TPM.0000502820.55789.3a](https://doi.org/10.1097/01.TPM.0000502820.55789.3a)]
71. Yadav D, Malik P, Dabas K, Singh P. Feedpal: understanding opportunities for chatbots in breastfeeding education of women in India. *Proc ACM Hum Comput Interact* 2019 Nov 07;3(CSCW):1-30. [doi: [10.1145/3359272](https://doi.org/10.1145/3359272)]
72. Gupta V, Arora N, Jain Y, Mokashi S, Panda C. Assessment on adoption behavior of first-time mothers on the usage of chatbots for breastfeeding consultation. *J Mahatma Gandhi Univ Med Sci Technol* 2021 Aug;6(2):64-68. [doi: [10.5005/jp-journals-10057-0161](https://doi.org/10.5005/jp-journals-10057-0161)]
73. Bennett V. Could artificial intelligence assist mothers with breastfeeding? *Br J Midwifery* 2018 Apr 02;26(4):212-213. [doi: [10.12968/bjom.2018.26.4.212](https://doi.org/10.12968/bjom.2018.26.4.212)]
74. Thirunavukarasu AJ, Ting DS, Elangovan K, Gutierrez L, Tan TF, Ting DS. Large language models in medicine. *Nat Med* 2023 Aug;29(8):1930-1940. [doi: [10.1038/s41591-023-02448-8](https://doi.org/10.1038/s41591-023-02448-8)] [Medline: [37460753](https://pubmed.ncbi.nlm.nih.gov/37460753/)]
75. de Souza J, Chamberlain K, Gupta S, Gao Y, Alshurafa N, Wang EJ. Opportunities in designing HCI tools for lactation consulting professionals. In: *Proceedings of the Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems*. 2022 Apr 22 Presented at: CHI EA '22; April 29-May 5, 2022; New Orleans, LA URL: <https://dl.acm.org/doi/10.1145/3491101.3519762> [doi: [10.1145/3491101.3519762](https://doi.org/10.1145/3491101.3519762)]

Abbreviations

- AI:** artificial intelligence
AUC: area under the curve
CNN: convolutional neural network
IBCLC: international board-certified lactation consultant
LC: lactation consultant
RGB: red, green, and blue
ROC-AUC: receiver operating characteristic area under the curve
VGG16: Visual Geometry Group model with 16 layers

Edited by K El Emam, B Malin; submitted 22.11.23; peer-reviewed by Z Li, L Juwara, J Li; comments to author 07.02.24; revised version received 20.04.24; accepted 09.05.24; published 24.06.24.

Please cite as:

De Souza J, Viswanath VK, Echterhoff JM, Chamberlain K, Wang EJ

Augmenting Telepostpartum Care With Vision-Based Detection of Breastfeeding-Related Conditions: Algorithm Development and Validation

JMIR AI 2024;3:e54798

URL: <https://ai.jmir.org/2024/1/e54798>

doi: [10.2196/54798](https://doi.org/10.2196/54798)

PMID:

©Jessica De Souza, Varun Kumar Viswanath, Jessica Maria Echterhoff, Kristina Chamberlain, Edward Jay Wang. Originally published in *JMIR AI* (<https://ai.jmir.org>), 24.06.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in *JMIR AI*, is properly cited. The complete bibliographic information, a link to the original publication on <https://www.ai.jmir.org/>, as well as this copyright and license information must be included.

Original Paper

Enhancing Type 2 Diabetes Treatment Decisions With Interpretable Machine Learning Models for Predicting Hemoglobin A1c Changes: Machine Learning Model Development

Hisashi Kurasawa^{1,2*}, PhD; Kayo Waki^{2*}, MPH, MD, PhD; Tomohisa Seki², MD, PhD; Akihiro Chiba^{1,3}, PhD; Akinori Fujino¹, PhD; Katsuyoshi Hayashi¹, PhD; Eri Nakahara^{1,2}, ME; Tsuneyuki Haga^{1,4}, PhD; Takashi Noguchi⁵, MD, PhD; Kazuhiko Ohe², MD, PhD

¹Nippon Telegraph and Telephone Corporation, Tokyo, Japan

²The University of Tokyo Hospital, Tokyo, Japan

³NTT DOCOMO, Inc, Tokyo, Japan

⁴NTT-AT IPS Corporation, Kanagawa, Japan

⁵National Center for Child Health and Development, Tokyo, Japan

*these authors contributed equally

Corresponding Author:

Kayo Waki, MPH, MD, PhD

The University of Tokyo Hospital

7-3-1 Hongo, Bunkyo-ku

Tokyo, 113-8655

Japan

Phone: 81 358006427

Email: kwaki-tyk@m.u-tokyo.ac.jp

Abstract

Background: Type 2 diabetes (T2D) is a significant global health challenge. Physicians need to assess whether future glycemic control will be poor on the current trajectory of usual care and usual-care treatment intensifications so that they can consider taking extra treatment measures to prevent poor outcomes. Predicting poor glycemic control from trends in hemoglobin A_{1c} (HbA_{1c}) levels is difficult due to the influence of seasonal fluctuations and other factors.

Objective: We sought to develop a model that accurately predicts poor glycemic control among patients with T2D receiving usual care.

Methods: Our machine learning model predicts poor glycemic control (HbA_{1c} ≥ 8%) using the transformer architecture, incorporating an attention mechanism to process irregularly spaced HbA_{1c} time series and quantify temporal relationships of past HbA_{1c} levels at each time point. We assessed the model using HbA_{1c} levels from 7787 patients with T2D seeing specialist physicians at the University of Tokyo Hospital. The training data include instances of poor glycemic control occurring during usual care with usual-care treatment intensifications. We compared prediction accuracy, assessed with the area under the receiver operating characteristic curve, the area under the precision-recall curve, and the accuracy rate, to that of LightGBM.

Results: The area under the receiver operating characteristic curve, the area under the precision-recall curve, and the accuracy rate (95% confidence limits) of the proposed model were 0.925 (95% CI 0.923-0.928), 0.864 (95% CI 0.852-0.875), and 0.864 (95% CI 0.86-0.869), respectively. The proposed model achieved high prediction accuracy comparable to or surpassing LightGBM's performance. The model prioritized the most recent HbA_{1c} levels for predictions. Older HbA_{1c} levels in patients with poor glycemic control were slightly more influential in predictions compared to patients with good glycemic control.

Conclusions: The proposed model accurately predicts poor glycemic control for patients with T2D receiving usual care, including patients receiving usual-care treatment intensifications, allowing physicians to identify cases warranting extraordinary treatment intensifications. If used by a nonspecialist, the model's indication of likely future poor glycemic control may warrant a referral to a specialist. Future efforts could incorporate diverse and large-scale clinical data for improved accuracy.

(JMIR AI 2024;3:e56700) doi:[10.2196/56700](https://doi.org/10.2196/56700)

KEYWORDS

AI; artificial intelligence; attention weight; type 2 diabetes; blood glucose control; machine learning; transformer

Introduction

Type 2 diabetes (T2D) affects an estimated 529 million people globally [1]. Hemoglobin A_{1c} (HbA_{1c}) serves as an indicator of poor glycemic control, reflecting the average blood glucose levels over 1 to 2 months. An increase in HbA_{1c} of 1 percentage point worsens cardiovascular disease risk by 1.2 times and mortality risk by 1.14 times [2]. According to the American Diabetes Association *Standards of Care in Diabetes* [3], target HbA_{1c} levels are set at 7% for many adults who are nonpregnant and 8% for patients with limited life expectancy or where the harms of treatment are greater than the benefits.

Physicians need to identify early signs of impending poor glycemic control in patients with T2D and act early to intensify treatment, via a combination of pharmacological and lifestyle interventions, to avoid poor outcomes. There are costs to intensified treatment, including side effects, so it is prudent to delay intensification until it is warranted by disease progression. Factors associated with poor glycemic control include age, duration of T2D treatment, treatment, race or ethnicity, and family history [4-7]. External factors such as seasonal variations affecting HbA_{1c} levels [8] complicate accurate glycemic control prediction.

People with T2D receive care from primary care physicians, not T2D specialists, in many areas including the United States, Europe [9], and Japan [10]. For example, two-thirds of people with T2D in Japan receive care from primary care physicians [10]. These nonspecialists may struggle to predict a patient's glycemic control. In Japan, approximately 60% of surveyed patients with T2D treated by nonspecialists experienced poor glycemic control (HbA_{1c}≥8%), with around 30% seeing worsened levels the following year, according to a survey on T2D treatment practices by primary care physicians [10].

Physicians regularly adjust a T2D patient's treatment, intensifying treatment when the clinical indications lead them to predict poor glycemic control. Despite this usual care, including treatment intensification, some patients still experience poor glycemic control. From 2015 to 2018, a total of 49.5% of US community-dwelling adults with diabetes had HbA_{1c}≥7% and 24.6% had HbA_{1c}≥8% [11]. A tool predicting poor glycemic control while under usual care, including usual-care treatment intensifications, could enhance treatment outcomes. It could alert physicians early enough to enable intensified modification of treatment, improving treatment outcomes for patients and increasing referrals to specialists when warranted by disease progression.

Machine learning (ML) has demonstrated success in predicting patient symptoms, including forecasting the onset of T2D [12] and predicting complications [13], and it is a promising approach to predicting poor glycemic control, although to our knowledge it has not previously been applied to this task. Glycemic control data are in general irregularly spaced, reflecting the variability in patient care appointment dates, with updates to outpatient

electronic health records (EHRs) occurring before and after clinical visits. Irregularly spaced data require preprocessing techniques such as interpolation, denoising autoencoders, and self-supervised learning [14-17]. Processing data with irregular intervals may hurt predictive performance [18], requiring careful consideration in developing artificial intelligence models.

Although ML models may provide good prediction performance, they often operate as "black boxes," with opaque reasoning and associated poor interpretability that makes it difficult for both physicians and patients to understand the logical process guiding decision-making [19]. To allow the interpretation of ML models, so that they are more acceptable to physicians [20,21] and patients [22], explainable artificial intelligence (XAI) has been studied [23]. It attempts to clarify temporal relationships of symptoms at each time point toward temporal interpretability based on patient trajectories [24,25], and this has been actively researched in the computer science field [26].

Since its introduction in 2017, the transformer model has excelled in various time-series predictive tasks, solidifying its position as a core technology across multiple fields [27-32]. The transformer model incorporates an attention mechanism simplifying the extraction of temporal relationships and setting it apart from other models [33-35]. The attention mechanism allows a model to selectively focus on different data points in the input sequence, assigning varying degrees of importance to each data point. Applied to the problem of predicting poor glycemic control, the attention mechanism can process irregularly spaced HbA_{1c} time series and quantify temporal relationships of past HbA_{1c} levels at each time point, following a model-specific approach in XAI [36].

This study aims to develop an ML tool that accurately and interpretably predicts poor glycemic control (HbA_{1c}≥8%) using irregularly spaced HbA_{1c} levels over the past year, in support of preventing T2D complications by enabling timely intensification of treatment. Although the treatment guidelines generally target an HbA_{1c} level of 7% or lower [3], higher levels are common in diabetes patients. In our clinical experience, levels of 8% and higher are a cause of great concern and trigger more intensive intervention. Accordingly, we have set 8% HbA_{1c} as the threshold for defining poor glycemic control.

Given the absence of prior studies in this specific area, we set target accuracy to be the receiver operating characteristic (ROC) area under the curve (AUC)>0.9 and precision-recall (PR)-AUC>0.8 based on our clinical endocrinology experience with diabetes treatment. These values are commonly used as a benchmark for good prediction accuracy in the ML field [37] and are consistent with the ROC-AUCs of past diabetes-related ML tasks ranging from 0.819 to 0.934 [38-42].

Drawing on our team's prior work in self-management support for T2D treatment [43] and predicting treatment discontinuations [44,45], we designed this task with the hope of overcoming barriers to implementing ML in clinical practice, believing it could significantly advance T2D diagnosis and treatment.

We hypothesize that an ML model can predict poor glycemic control in patients with T2D under usual care. Our specific research question is whether a transformer-based model, incorporating temporal relationships of HbA_{1c} levels, can accurately and interpretably predict instances of poor glycemic control (HbA_{1c}≥8%). Our approach is novel in how it overcomes challenges posed by irregularly spaced HbA_{1c} time series.

Methods

Data Sets and Preprocessing

All data were collected from EHRs at the University of Tokyo Hospital, which included 7787 patients who visited the hospital and had diagnostic codes indicative of T2D. The data were recorded in the EHRs between January 1, 2006, and December 31, 2015. The data, including treatment decisions and outcomes, were reflective of care by T2D specialists. Only HbA_{1c} levels were used in the ML model.

ML Models

Given the irregularly spaced data, we organized the data into Monday-to-Sunday weeks and quantized the data to a single value per week, using the average in the case of multiple measurements and treating weeks with no values as having missing values [46]. This approach allowed the ML model to treat irregularly spaced data spanning N years as regularly spaced data consisting of $N \times 365/7$ (rounded up to the nearest integer) values, that is, we treat all data as weekly data. We did not perform preprocessing, including interpolation, on missing values in the regularly spaced data. No normalization, outlier removal, or dimensionality reduction were performed on the HbA_{1c} levels. Typical ML models such as LightGBM address

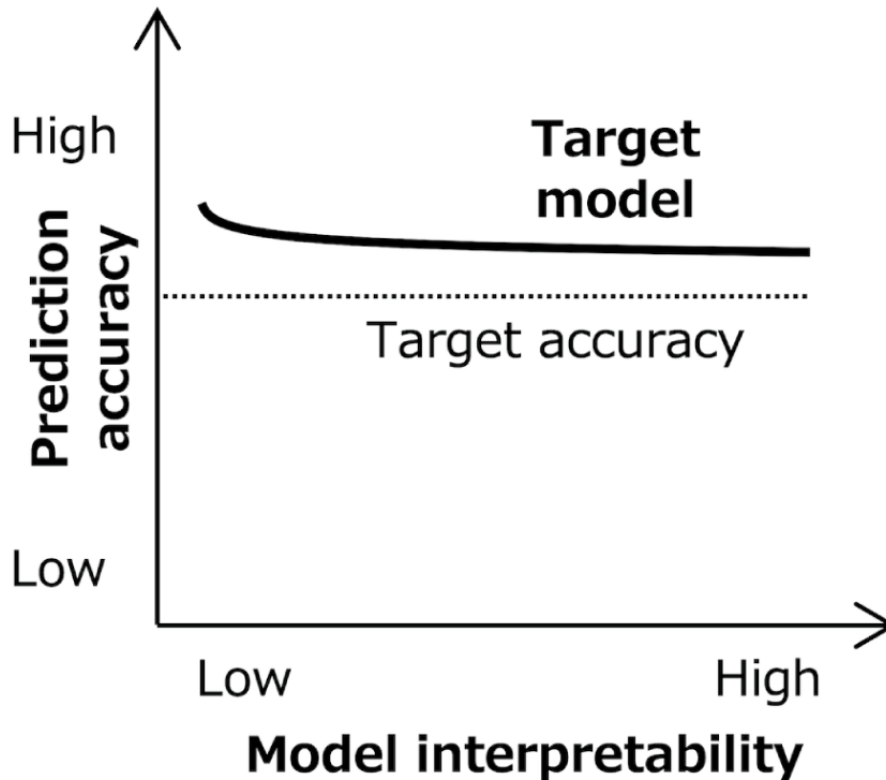
missing values via interpolation or replacement before learning. In contrast, we adopted an approach that ignores and skips missing values.

We designed a transformer model (Table 1) that takes as input an irregularly spaced time series of HbA_{1c} levels spanning over the past year or more and outputs a binary assessment of poor glycemic control (HbA_{1c}≥8%) within the subsequent year. The model incorporates 2 types of attention layers: self-attention, designed to extract temporal relationships from past irregularly spaced HbA_{1c} levels, and cross-attention, used to predict poor glycemic control based on these temporal relationships. The self-attention weights are optimized through self-supervised learning. This involves the task of predicting the next HbA_{1c} level using a time series of weekly spaced past HbA_{1c} levels with missing values, where the past levels are used as both input and output. We use an attention mask mechanism that completely ignores missing values by setting their self-attention weight to 0, allowing us to learn using values with irregular spacing due to missing values as is. This is similar to the process of padding in language models. The cross-attention weights are optimized through supervised learning. This involves the task of predicting the class representing the likelihood of future poor glycemic control using the latent variables transformed by self-attention from the past HbA_{1c} time series. We used causal masking in both learning tasks to prevent the model from referencing future data, ensuring that the model makes predictions considering the causal relationship between past symptoms and future symptoms. Conceptually, given that we constrained the model to improve interpretability, we expect a slightly lower prediction accuracy than that of an unconstrained model (Figure 1), and 1 goal is to minimize this interpretability penalty.

Table 1. Model details (transformer architecture).

Configure	Value
Encoder layers including self-attention blocks, n	4
Decoder layers including cross-attention blocks, n	4
Heads in the attention, n	4
Transformer hidden size, n	128
Transformer feedforward neural network hidden size, n	512
Optimizer method	Adam
Loss function	Focal Loss
Learning rate	1×10^{-4}
Batch size, n	512
Iterations, n	20,000 (no early stopping)
Library	Python (version 3.11) and PyTorch (version 2.2.0)

Figure 1. Conceptual trade-off between prediction accuracy and interpretability, for a given level of computational complexity.



This model makes a yes or no decision on future poor glycemic control, with a threshold as a free variable allowing a tradeoff among true or false negatives or positives. We set the threshold to maximize the F_1 -score (the harmonic mean of the ROC and PR values) using training data, resulting in a tool that made a binary prediction as to whether or not, in the next year, glycemic control will be poor. The training data include treatment intensification by specialists making their own assessments of likely future glycemic control. As such, the predicted poor glycemic control occurs despite any usual-care intensification of treatment prescribed by the attending specialist physicians. In other words, a prediction of poor glycemic control indicates a case likely warrants special attention and intervention, as usual-care intensification of treatment is predicted to be insufficient.

Temporal Data Usage

Our analysis sought to determine the length of the HbA_{1c} time series needed to achieve the target accuracy. Training and testing were separated by period using the well-established time series prediction accuracy evaluation method [47]. We used as a reference the date on which a patient took an HbA_{1c} test in 2013. We used the HbA_{1c} time series for the N years before the reference date as training input and the occurrence or absence of poor glycemic control (HbA_{1c} ≥ 8%) within 1 year from the reference date as the training output. Then, we tested the resulting model using the same procedure, but for the following year, 2014, selecting an appropriate choice for N, the length of training data. We evaluated the predictive performance of the resulting model using 7 years of test data, sliding the reference dates from 2007 to 2013, using the rolling-origin procedure [47].

The training input or output period and the testing output period do not overlap, and therefore there was no leakage into predictive evaluation. Data for a given patient will in general have some time samples in the training data and some in the test data, but since patient identification is not an input to the model, the model does not identify specific patients.

Statistical Methods

We analyzed the characteristics of patients in the data set using means, SDs, and frequency counts. We performed all statistical analyses using custom Python code. We used the *Python* (version 3.11) and *PyTorch* (version 2.2) libraries for developing the transformer model, the *Numpy* (version 1.26) and *Pandas* (version 2.2) libraries for managing data sets, and the *scikit-learn* (version 1.4) library for evaluating predictive accuracy.

We compared our model with an established ML method recognized for high accuracy. There were no studies directly addressing our task, but validations on similar T2D prediction tasks favored LightGBM [48,49], making it our chosen reference for comparisons. While LightGBM is acknowledged for its superior predictive performance, it is not inherently interpretable. The model's complexity and intricate decision tree paths make it difficult to provide a straightforward interpretation of its predictions. Our reference LightGBM model takes as input equally spaced HbA_{1c} data and outputs a binary assessment of poor glycemic control (HbA_{1c} ≥ 8%).

We compared the transformer model and LightGBM using the evaluation metrics of ROC-AUC, PR-AUC, accuracy rate, and F_1 -score, with 95% CI using the bootstrap method.

Ethical Considerations

This study was approved by the Institutional Review Board of the University of Tokyo School of Medicine (10705-(3)) and was conducted per the Declaration of Helsinki. This was a retrospective, noninterventional database study without patient involvement. Confidentiality was safeguarded by the University of Tokyo Hospital. According to the Guidelines for Epidemiological Studies of the Ministry of Health, Labour and Welfare of Japan, written informed consent was not required. Information about this study was available to patients on a website, and patients have the right to cease registration of their data at any time [50].

Results

Patient Data

We analyzed 7787 patients (Table 2). Although specialist physicians were providing usual care and prescribing treatment intensifications based on their clinical judgment, 57.83% (n=4504) of patients had an HbA_{1c} over 8% at least once. The number of HbA_{1c} tests per year was 7.7 (SD 2.8). In other words, the missingness level of weekly spaced past HbA_{1c} levels for a year was $1 - 7.7 / \text{ROUNDUP}(365/7) = 85.5\%$. The age group with the highest number of individuals is the aged 70-80 years category, comprising 2347 people, accounting for 30.14% of the patients. In addition to diabetes, more than 45% of patients had diseases such as essential (primary) hypertension, hypertensive heart disease, pure hypercholesterolemia, and astigmatism. Each patient had multiple records, leading to 323,825 records used in our analysis.

Table 2. Characteristics of patients.

Characteristics	Records (n=323,825)	Patients (n=7787)
Feature used in the model		
HbA_{1c}^a		
Mean (SD)	7.1 (1.1)	— ^b
<6%, n (%)	42,495 (13.12)	4103 (52.69)
6%-7%, n (%)	137,968 (42.61)	6666 (85.6)
7%-8%, n (%)	89,875 (27.75)	5770 (74.1)
≥8%, n (%)	53,487 (16.52)	4504 (57.84)
Tests per year, mean (SD)	7.7 (2.8)	—
Features not used in the model		
Gender		
Male, n (%)	193,976 (59.9)	4726 (60.7)
Female, n (%)	129,849 (40.1)	3061 (39.3)
Age (years)		
Mean (SD)	—	67.5 (13.6)
10-20, n (%)	—	1 (0.01)
20-30, n (%)	—	58 (0.74)
30-40, n (%)	—	255 (3.27)
40-50, n (%)	—	585 (7.51)
50-60, n (%)	—	1006 (12.92)
60-70, n (%)	—	2058 (26.43)
70-80, n (%)	—	2347 (30.14)
80-90, n (%)	—	1322 (16.98)
90-100, n (%)	—	149 (1.91)
100-110, n (%)	—	6 (0.08)
Top 10 most common diseases		
E14: unspecified diabetes mellitus, n (%)	—	5495 (70.57)
I10: essential (primary) hypertension, n (%)	—	5023 (64.5)
E11: hypertensive heart disease, n (%)	—	3715 (47.71)
E780: pure hypercholesterolemia, n (%)	—	3661 (47.01)
H522: astigmatism, n (%)	—	3636 (46.69)
E785: hyperlipidemia, unspecified, n (%)	—	3490 (44.82)
K590: constipation, n (%)	—	3353 (43.06)
K210: gastro-esophageal reflux disease with esophagitis, n (%)	—	2937 (37.72)
K295: chronic gastritis, unspecified, n (%)	—	2756 (35.39)
Top 10 most common medicines		
Metformin hydrochloride, n (%)	—	2541 (32.63)
Sitagliptin phosphate hydrate, n (%)	—	2177 (27.96)
Glimepiride, n (%)	—	2036 (26.15)
Pioglitazone hydrochloride, n (%)	—	1641 (21.07)
Insulin glargine (genetical recombination), n (%)	—	1597 (20.51)
Rosuvastatin calcium, n (%)	—	1458 (18.72)

Characteristics	Records (n=323,825)	Patients (n=7787)
Voglibose, n (%)	—	1430 (18.36)
Atorvastatin calcium hydrate, n (%)	—	1323 (16.99)
Insulin aspart (genetical recombination), n (%)	—	1277 (16.4)
Vildagliptin, n (%)	—	1187 (15.24)

^aHbA_{1c}: hemoglobin A_{1c}.

^bNot applicable.

Prediction Performance for HbA_{1c} Time Series Lengths

We assessed using different lengths of past HbA_{1c} time series (Table 3) as both training and test inputs to the model to determine the most effective period for predicting poor glycemic control. Extending the input period beyond 1 year did not yield a statistically significant difference within a 95% CI (Figures

2 and 3). This study's objectives of achieving ROC-AUC>0.9 and PR-AUC>0.8 were attainable with just 1 year of past HbA_{1c} time series. Comparing prediction accuracy with LightGBM revealed no significant differences within the 95% CI, indicating nearly equivalent performance between the transformer and LightGBM. As a result, we settled on a final model that is based on using 1 year of prior data for training.

Table 3. Test data set size for the evaluation of various hemoglobin A_{1c} (HbA_{1c}) time series lengths.

Length of past HbA _{1c} time series	Records (R), n	Records with poor glycemic control (T), n	T/R, %	Patients, n	Records per patient, mean (SD)	Weekly spaced data with values in input data, mean (SD)
1	25,564	6818	26.7	4661	5.5 (2.7)	7.3 (2.7)
2	25,594	6827	26.7	4672	5.5 (2.7)	13.2 (5.6)
3	25,611	6831	26.7	4676	5.5 (2.7)	18.8 (8.6)
4	25,618	6831	26.7	4678	5.5 (2.7)	24.1 (11.8)
5	25,621	6832	26.7	4678	5.5 (2.7)	28.9 (15.1)

Figure 2. Predictive performance using ROC-AUC as a measure for various HbA_{1c} time series lengths using test data reference dates in 2014. HbA_{1c}: hemoglobin A_{1c}; ROC-AUC: area under the receiver operating characteristic curve.

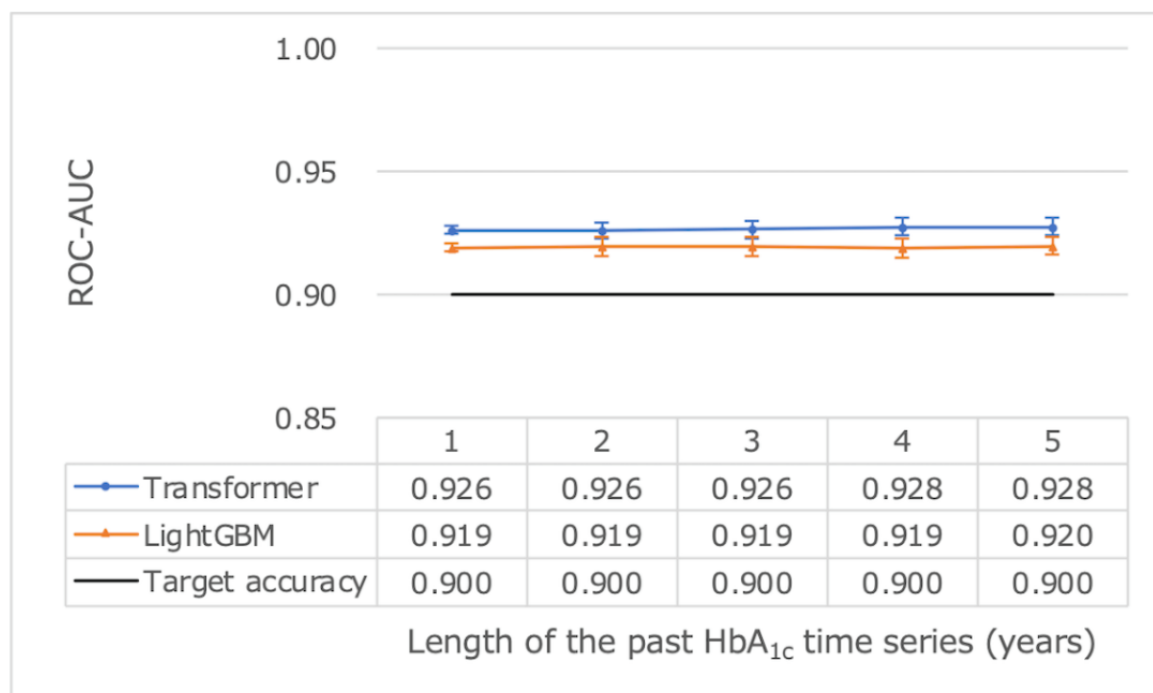
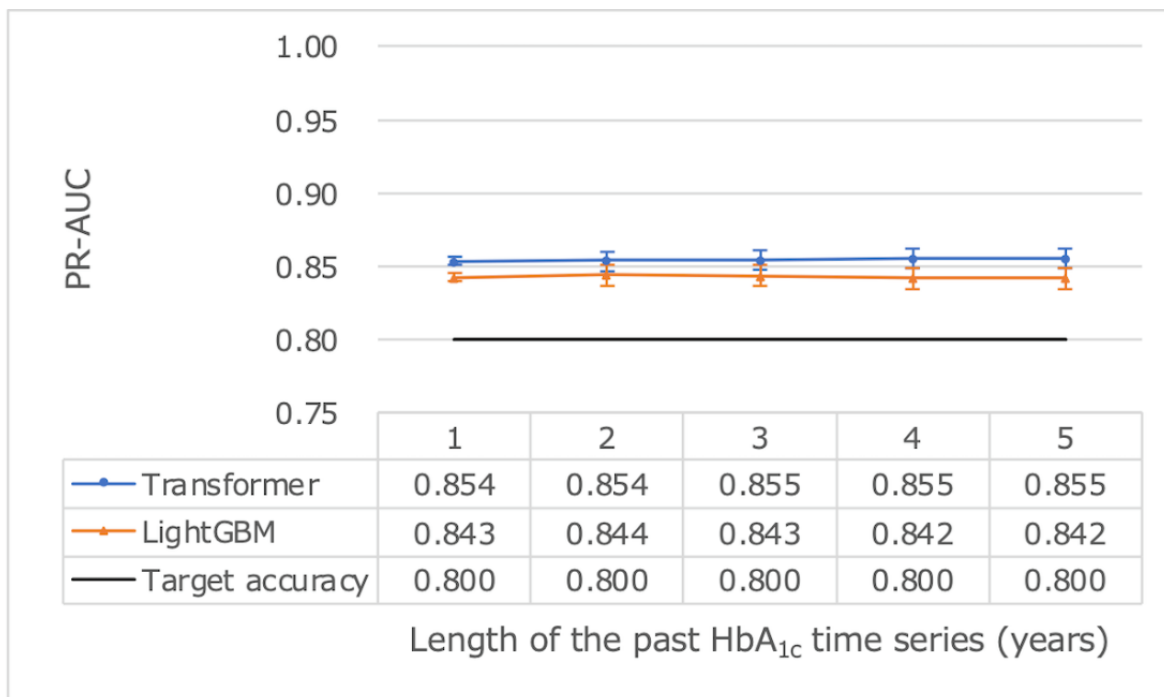


Figure 3. Predictive performance using PR-AUC as a measure for various HbA_{1c} time series lengths using test data reference dates in 2014. HbA_{1c}: hemoglobin A_{1c}; PR-AUC: area under the precision-recall curve.



Prediction Performance Over the Full Data Set

We assessed whether the resulting model, using 1 year of prior data for training, could consistently achieve the target accuracy over the available 7 years of test data (Table 4). Despite some fluctuation in prediction accuracy, the target was achieved over the entire 7-year period (Figures 4 and 5). The ROC-AUC (95% confidence limits) for transformer was 0.925 (95% CI

0.923-0.928; Figure 6), compared to LightGBM’s 0.920 (95% CI 0.918-0.923), and the PR-AUC (95% confidence limits) for transformer was 0.864 (95% CI 0.852-0.875; Figure 7), compared to LightGBM’s 0.857 (95% CI 0.846-0.868). The average accuracy rate (95% confidence limits) for the transformer was 0.864 (95% CI 0.860-0.869), comparable to LightGBM’s 0.861 (95% CI 0.857-0.865).

Table 4. Test data set size for the evaluation of various hemoglobin A_{1c} (HbA_{1c}) time series lengths.

Year of the test data	Records (R), n	Records with poor glycemic control (T), n	T/R, %	Patients, n	Records per patient, mean (SD)	Weekly spaced data with values in input data, mean (SD)
2007	22,520	7176	31.9	3221	7 (3.1)	8 (2.9)
2008	24,775	7517	30.3	3626	6.8 (3.1)	8.1 (2.9)
2009	26,144	8444	32.3	2973	6.6 (3)	8 (2.9)
2010	27,124	8521	31.4	4260	6.4 (3)	7.8 (2.9)
2011	26,661	7687	28.8	4377	6.1 (3)	7.7 (2.8)
2012	26,259	6944	26.4	4412	6 (2.9)	7.5 (2.7)
2013	25,945	7281	28.1	4533	5.7 (2.8)	7.4 (2.7)
2014	25,564	6818	26.7	4661	5.5 (2.7)	7.3 (2.7)

Figure 4. Predictive performance over time using ROC-AUC as a measure using test data reference dates ranging from 2008 to 2014. ROC-AUC: area under the receiver operating characteristic curve.

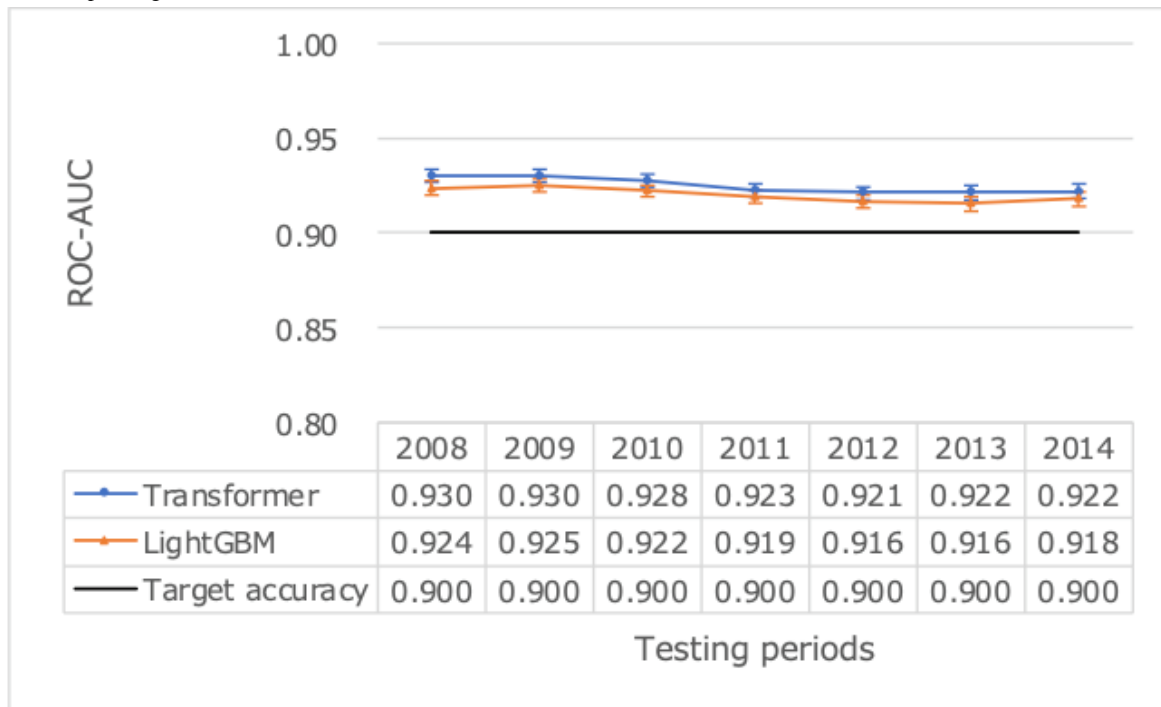


Figure 5. Predictive performance over time using PR-AUC as a measure using test data reference dates ranging from 2008 to 2014. PR-AUC: area under the precision-recall curve.

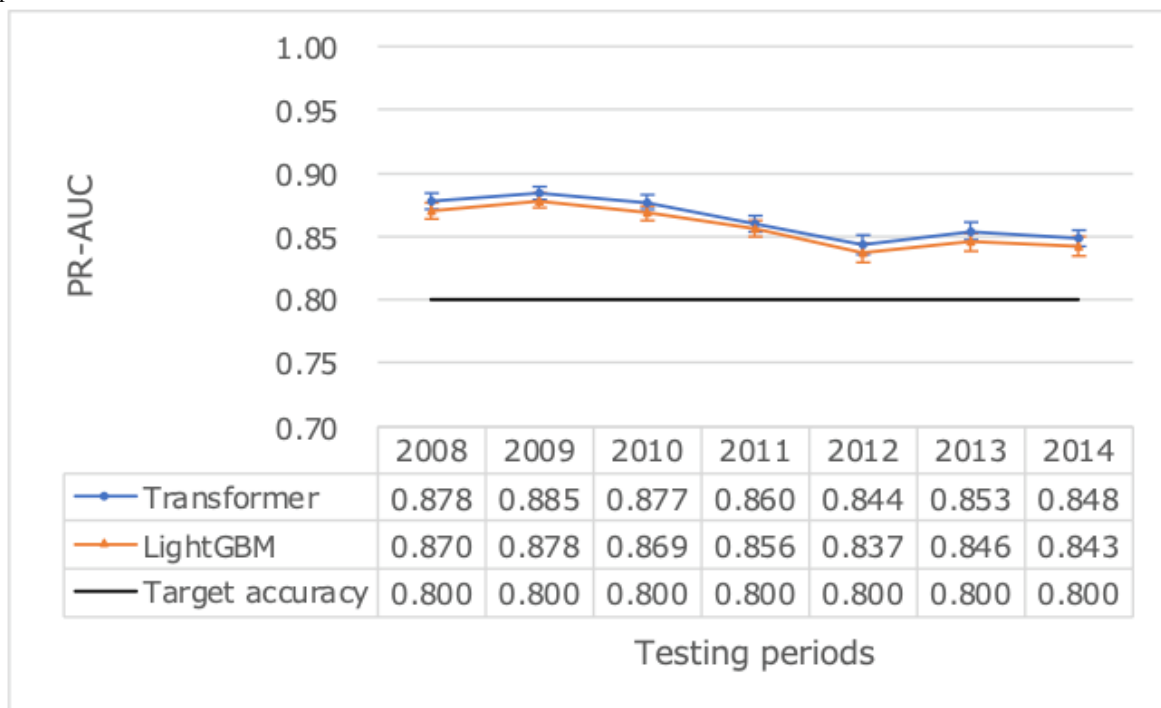


Figure 6. Predictive performance over time using ROC curve as a measure using test data reference dates ranging from 2008 to 2014. AUC: area under the curve; FPR: false positive rate; HbA_{1c}: hemoglobin A_{1c}; ROC: receiver operating characteristic; TPR: true positive rate.

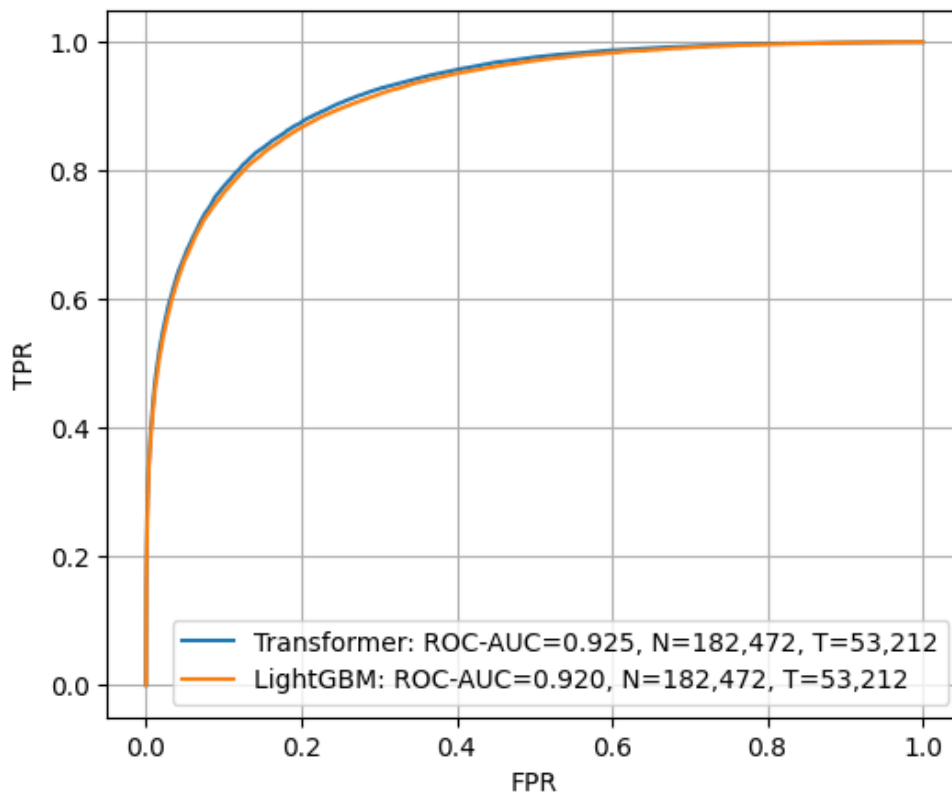
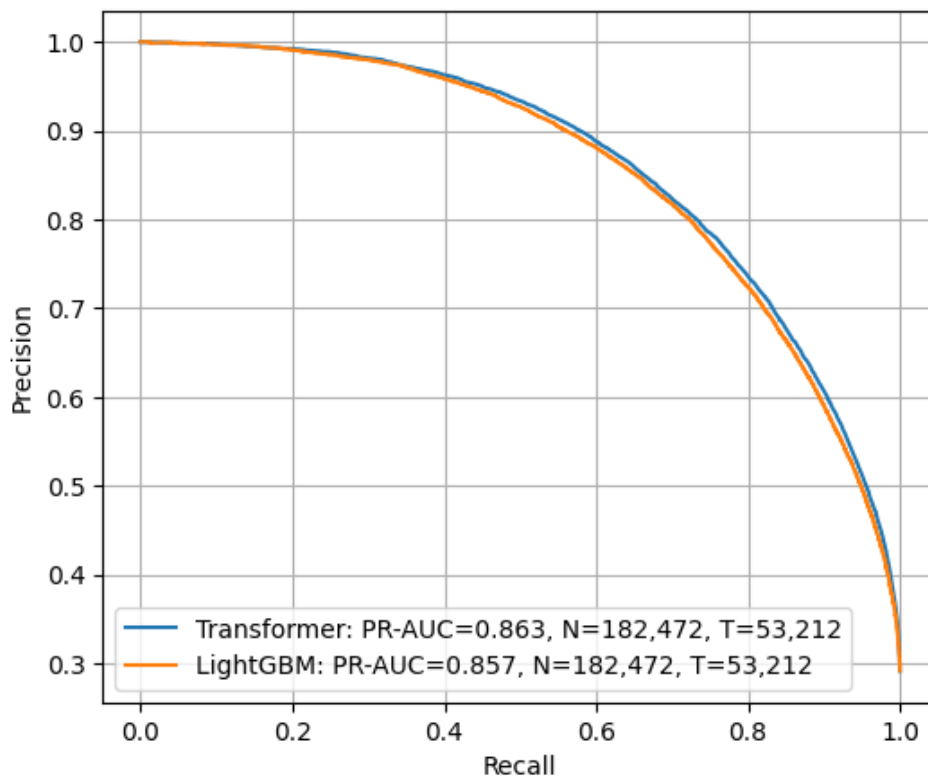


Figure 7. Predictive performance over time using PR curve as a measure using test data reference dates ranging from 2008 to 2014. AUC: area under the curve; HbA_{1c}: hemoglobin A_{1c}; PR: precision-recall.



Interpretability

The proposed model extracts temporal relationships from past irregularly spaced HbA_{1c} levels using self-attention and determines the contribution of each HbA_{1c} level to the prediction of glycemic control using cross-attention. An example of the extracted results is shown in Figure 8.

Figures 9-11 plot the average values of HbA_{1c} levels, self-attention weights, and cross-attention weights for 4 groups: true positives with transformer and true positives with LightGBM, true negatives with transformer and true negatives with LightGBM, true positives with transformer and false negatives with LightGBM, and false negatives with transformer and true positives with LightGBM. The group with true positive

results in both models had an average HbA_{1c} level of 8% or higher, whereas the group with true negative results in both models had an average HbA_{1c} level of less than 7%. The weight of older self-attention was larger in the former group, and the weight of recent cross-attention was smaller in the latter group. The group containing true positives by transformer and false negatives with LightGBM had an average HbA_{1c} level of around 7.5%, had a smaller recent self-attention weight than the other groups, and had a similar trend of cross-attention weights as the group of true negatives with both models. The group that was false negative with transformer and true positive with LightGBM tended for HbA_{1c} to fall from the 8% range to the 7% range, and both recent self-attention and cross-attention were greater than other groups.

Figure 8. Example of HbA_{1c} levels, self-attention weights, and cross-attention weights. HbA_{1c}: hemoglobin A_{1c}.

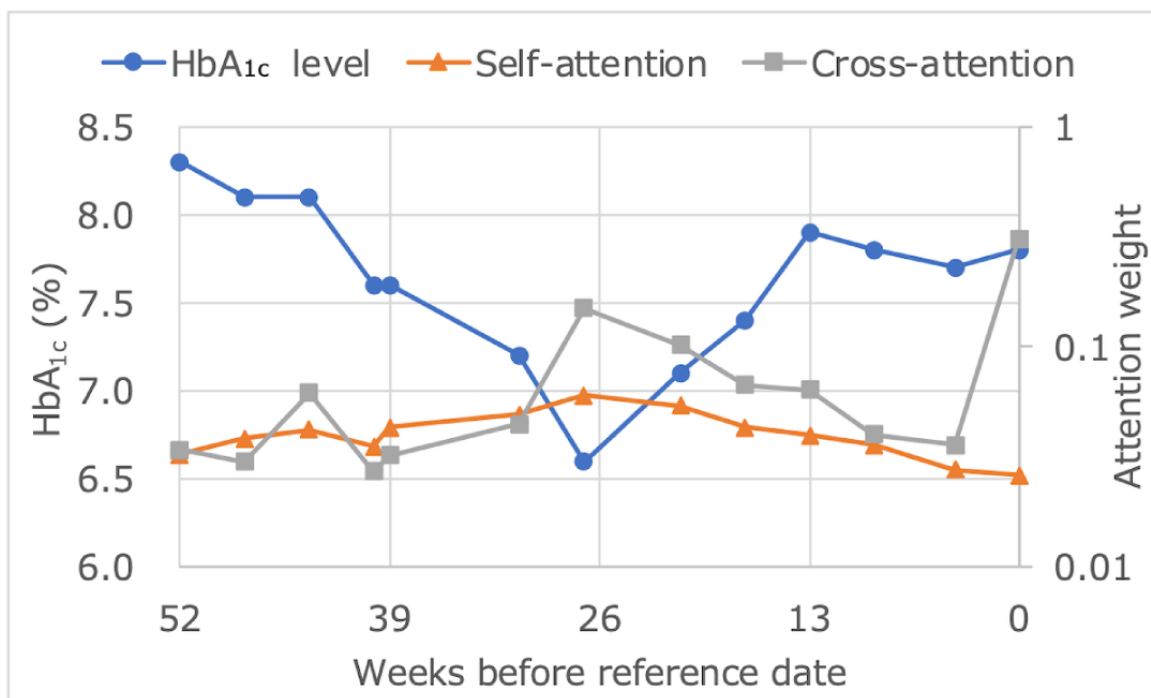


Figure 9. Average levels of HbA_{1c} time series. HbA_{1c}: hemoglobin A_{1c}.

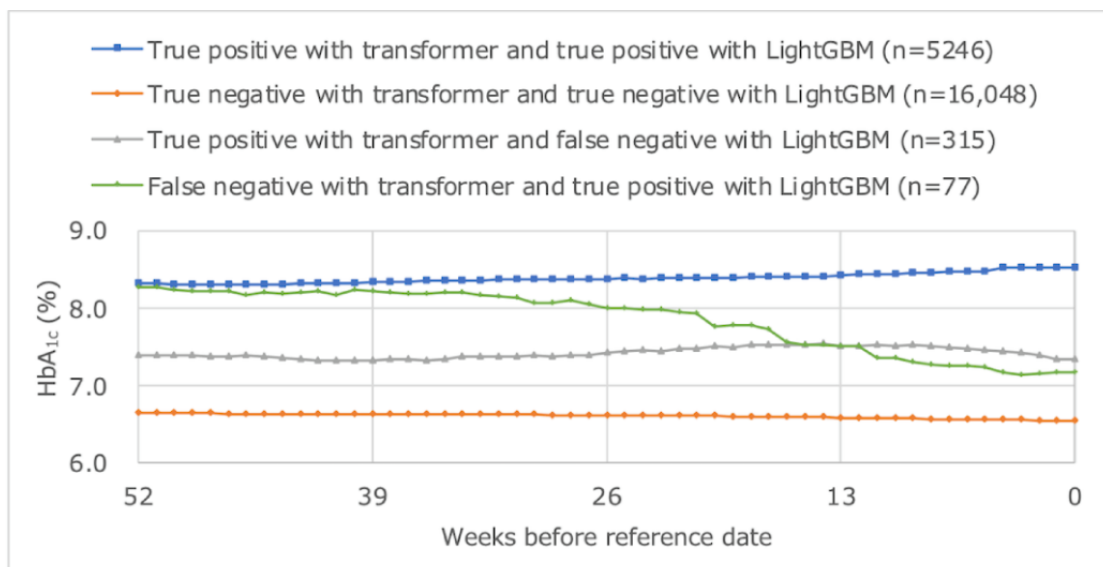
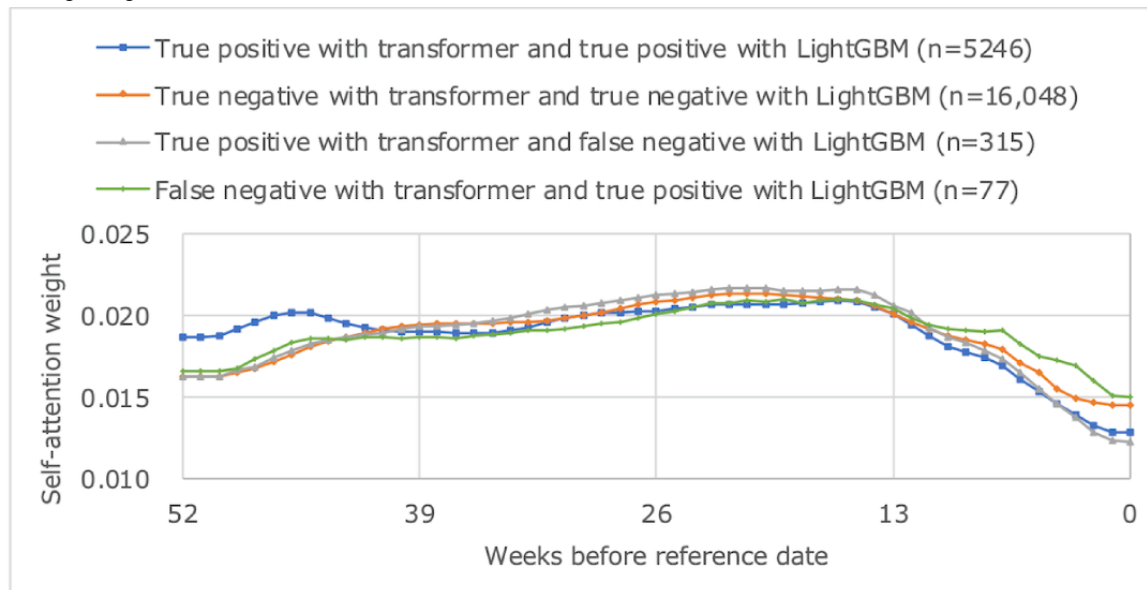
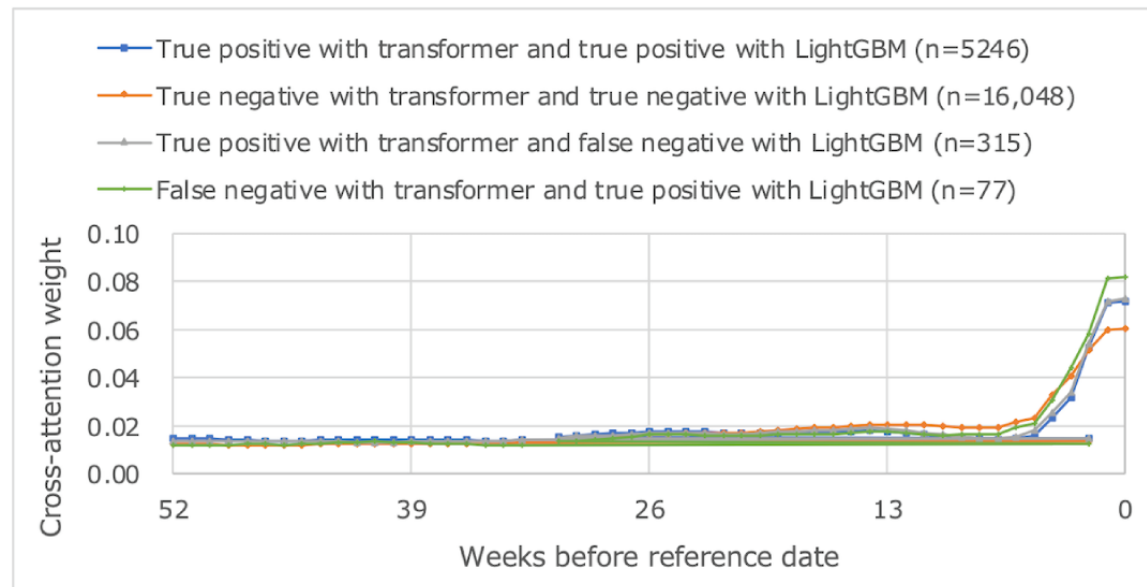


Figure 10. Average weight of self-attention.**Figure 11.** Average weight of cross-attention.

Discussion

Evaluation of the Predictive Accuracy

Our results show that, despite usual care by specialist physicians, poor glycaemic control was common, affecting 57.83% (4504/7787) of patients. By highlighting cases with a high likelihood of poor glycaemic control despite normal treatment intensifications, the proposed model provides new information to physicians, identifying patients who may benefit from extraordinary treatment intensification.

Balancing high predictive accuracy with interpretability is vital for acceptance by patients and physicians. The proposed model achieved impressive predictive accuracy, with ROC-AUC above 0.9, PR-AUC above 0.8, and an overall accuracy of 0.864. For physicians, ROC-AUC above 0.9 suggests excellent performance in distinguishing between patients who will have poor glycaemic control and patients who will have good glycaemic

control. Similarly, PR-AUC above 0.8 indicates excellent performance in providing accurate prediction while minimizing false positives. LightGBM, a widely respected model in ML, serves as a benchmark. The proposed model slightly surpassed the performance of LightGBM, implying that the proposed model can offer physicians a reliable tool for predicting poor glycaemic control.

Accuracy did not increase with longer training data lengths. The model achieved accurate predictions with just 1 year of training data, suggesting that recent glycaemic control plays a dominant role in prediction outcomes. However, the actual future glycaemic control is influenced by factors not accounted for in the current model, such as medications, exercise, diet, and other lifestyle factors.

While the proposed model demonstrated comparable predictive accuracy to LightGBM within this experiment's scope, further improvement may be possible with extensive training data.

Transformer models, known for power-law characteristics, benefit from scale-ups [51], and expanding this study to multiple hospitals could explore potential performance enhancements and test the applicability of the power-law in the medical field.

Interpretability

The cross-attention weights were very similar for the group that was true positive in both models and the group that was true positive in transformer and false negative in LightGBM. This suggests that the proposed model consistently made predictions by capturing sufficient features, while the benchmark LightGBM might have captured extraneous features. On the other hand, when the proposed model performed worse than LightGBM, as observed in the group of false negatives with the transformer and true positives with LightGBM, it appears that the cross-attention strongly responded to the decreasing trend in HbA_{1c}, leading to a prediction failure. These prediction failures accounted for only 0.30% (77/25,564) of cases.

Limitations

Our study has notable limitations. First, the data were sourced from past records at a single hospital, limiting generalizability. We have not confirmed prediction accuracy for new patients, as we used a rolling origin procedure. While we separated the data into training and testing sets based on time duration, some patients still overlap between these sets. While this approach is useful for assessing the model's performance within the hospital where it is trained, it poses challenges when applying a model trained in one hospital to another. The intensification of treatment may depend on factors specific to individual patients, the treatment strategies of individual physicians and hospitals, guidelines, and varying treatment trends across countries. Further work is needed to verify the extent to which the model needs to be customized for different environments.

Second, ML reflects majority characteristics, potentially limiting applicability to diverse patient populations. In the data set used in the experiment, as shown in Table 2, 40% of patients have 7 diseases, and patient characteristics are biased. Prediction failure analysis needs to be further scrutinized, including versus patient characteristics. We should examine this issue by comparing prediction accuracy for each patient cluster.

Third, the model uses only HbA_{1c} levels as inputs. We incorporated prescription and other laboratory tests as explanatory variables during preliminary validation, but both our proposed model and LightGBM did not show improved predictive accuracy. Future work should further explore incorporating clinical data beyond HbA_{1c}. EHRs contain patient history represented in categorical, numeric, text, and images that are still underused. We should devise model designs based

on cutting-edge multimodal modeling using the transformer [52-54].

Fourth, the interpretability of the model expresses temporal relationships numerically, lacking readability. To enhance clarity and visualization of the information that physicians require, it is essential to solidify the user interface or user experience concepts. There is a need for further consultation with physicians to determine an interface that would effectively communicate interpretability. Additionally, to increase the interpretability of this method, an approach that combines it with traditional XAI technologies [36] such as SHAP and LIME should be investigated.

Fifth, this was a backward-looking study, using past data, and the essential next phase is to assess the model's predictive capabilities in clinical practice. There is a need for a careful exploration of the model's effectiveness in real clinical scenarios.

Future Research Direction

Our ultimate goal is to improve the treatment outcomes of diabetes. Merely predicting poor glycemic control alone cannot achieve this goal. By providing predictive results to physicians and reinforcing treatment, we can demonstrate the value of the predictions. Future research could focus on improving predictions by incorporating additional clinical data beyond HbA_{1c} levels. Exploring the applicability of the model in diverse populations will help assess its generalizability and institution-specific variations. Implementing the model in clinical practice for real-time predictions, possibly through randomized controlled trials, would elucidate its impact on clinical decision-making and patient outcomes. Moreover, expanding the scope to predict the impact of treatment changes as well [55] could further enhance the model's utility in diabetes management.

Conclusions

The proposed model addresses the challenge of identifying patients with T2D who will have poor glycemic control, increasing the risk of complications, despite usual care by specialist physicians. The model achieves highly accurate predictions, with an accuracy of 0.864, and provides good interpretability from the irregularly spaced HbA_{1c} values commonly observed in clinical settings. The model balances desirable predictive accuracy and interpretability in clinical practice, enhancing the acceptability of ML. Future efforts should focus on further improving accuracy and interpretability by incorporating additional features beyond HbA_{1c} and validating large clinical data sets.

Acknowledgments

We thank Daniel Lane, ME, MBA, for his support in paper editing and scientific discussions. We made no use of generative artificial intelligence in the development of this paper. This work was funded by The University of Tokyo and Nippon Telegraph and Telephone Corporation (NTT) in a joint research program at the University of Tokyo Center of Innovation, Sustainable Life Care, and the Ageless Society dedicated to Self-Managing Healthcare in the Aging Society of Japan. The funding source had no role in the design and conduct of this study; collection, management, analysis, and interpretation of the data; preparation, review,

or approval of this paper; and decision to submit this paper for publication. The content is solely the responsibility of the authors and does not necessarily represent the official views of the University of Tokyo Center of Innovation.

Data Availability

The data in this study are not openly available because of the restrictions imposed by the research ethics committees that approved this study.

Conflicts of Interest

HK, KH, AF, and EN are employees of Nippon Telegraph and Telephone Corporation (NTT), Tokyo, Japan. AC was an employee of NTT, and he is now an employee of NTT DOCOMO, Inc, Tokyo, Japan. TH was an employee of NTT, and he is now the chief executive officer of NTT-AT IPS Corporation, Kanagawa, Japan. KW received research funding from NTT and the University of Tokyo Center of Innovation, Sustainable Life Care, and the Ageless Society dedicated to Self-Managing Healthcare in the Aging Society of Japan.

References

1. GBD 2021 Diabetes Collaborators. Global, regional, and national burden of diabetes from 1990 to 2021, with projections of prevalence to 2050: a systematic analysis for the Global Burden of Disease study 2021. *Lancet* 2023;402(10397):203-234 [FREE Full text] [doi: [10.1016/S0140-6736\(23\)01301-6](https://doi.org/10.1016/S0140-6736(23)01301-6)] [Medline: [37356446](https://pubmed.ncbi.nlm.nih.gov/37356446/)]
2. Chen YY, Lin YJ, Chong E, Chen PC, Chao TF, Chen SA, et al. The impact of diabetes mellitus and corresponding HbA1c levels on the future risks of cardiovascular disease and mortality: a representative cohort study in Taiwan. *PLoS One* 2015;10(4):e0123116 [FREE Full text] [doi: [10.1371/journal.pone.0123116](https://doi.org/10.1371/journal.pone.0123116)] [Medline: [25874454](https://pubmed.ncbi.nlm.nih.gov/25874454/)]
3. ElSayed NA, Aleppo G, Aroda VR, Bannuru RR, Brown FM, Bruemmer D, et al. 6. Glycemic targets: standards of care in diabetes-2023. *Diabetes Care* 2023;46(Suppl 1):S97-S110 [FREE Full text] [doi: [10.2337/dc23-S006](https://doi.org/10.2337/dc23-S006)] [Medline: [36507646](https://pubmed.ncbi.nlm.nih.gov/36507646/)]
4. Harris MI, Eastman RC, Cowie CC, Flegal KM, Eberhardt MS. Racial and ethnic differences in glycemic control of adults with type 2 diabetes. *Diabetes Care* 1999;22(3):403-408. [doi: [10.2337/diacare.22.3.403](https://doi.org/10.2337/diacare.22.3.403)] [Medline: [10097918](https://pubmed.ncbi.nlm.nih.gov/10097918/)]
5. Goudswaard AN, Stolk RP, Zuihthoff P, Rutten GEHM. Patient characteristics do not predict poor glycaemic control in type 2 diabetes patients treated in primary care. *Eur J Epidemiol* 2004;19(6):541-545. [doi: [10.1023/b:ejep.0000032351.42772.e7](https://doi.org/10.1023/b:ejep.0000032351.42772.e7)] [Medline: [15330126](https://pubmed.ncbi.nlm.nih.gov/15330126/)]
6. Haghghatpanah M, Nejad ASM, Haghghatpanah M, Thunga G, Mallayasamy S. Factors that correlate with poor glycemic control in type 2 diabetes mellitus patients with complications. *Osong Public Health Res Perspect* 2018;9(4):167-174 [FREE Full text] [doi: [10.24171/j.phrp.2018.9.4.05](https://doi.org/10.24171/j.phrp.2018.9.4.05)] [Medline: [30159222](https://pubmed.ncbi.nlm.nih.gov/30159222/)]
7. Juarez DT, Sentell T, Tokumaru S, Goo R, Davis JW, Mau MM. Factors associated with poor glycemic control or wide glycemic variability among diabetes patients in Hawaii, 2006-2009. *Prev Chronic Dis* 2012;9:120065 [FREE Full text] [doi: [10.5888/pcd9.120065](https://doi.org/10.5888/pcd9.120065)] [Medline: [23017247](https://pubmed.ncbi.nlm.nih.gov/23017247/)]
8. Gikas A, Sotiropoulos A, Pastromas V, Papazafiropoulou A, Apostolou O, Pappas S. Seasonal variation in fasting glucose and HbA1c in patients with type 2 diabetes. *Prim Care Diabetes* 2009;3(2):111-114. [doi: [10.1016/j.pcd.2009.05.004](https://doi.org/10.1016/j.pcd.2009.05.004)] [Medline: [19535310](https://pubmed.ncbi.nlm.nih.gov/19535310/)]
9. Davies MJ, Aroda VR, Collins BS, Gabbay RA, Green J, Maruthur NM, et al. Management of hyperglycaemia in type 2 diabetes, 2022. a consensus report by the American Diabetes Association (ADA) and the European Association for the Study of Diabetes (EASD). *Diabetologia* 2022;65(12):1925-1966 [FREE Full text] [doi: [10.1007/s00125-022-05787-2](https://doi.org/10.1007/s00125-022-05787-2)] [Medline: [36151309](https://pubmed.ncbi.nlm.nih.gov/36151309/)]
10. Primary care physicians' practices in diabetes treatment. URL: <https://www.jdome.jp/doc/jdome-2021-64jds.pdf> [accessed 2024-06-22]
11. American Diabetes Association Professional Practice Committee. 1. Improving care and promoting health in populations: standards of care in diabetes-2024. *Diabetes Care* 2024;47(Suppl 1):S11-S19. [doi: [10.2337/dc24-S001](https://doi.org/10.2337/dc24-S001)] [Medline: [38078573](https://pubmed.ncbi.nlm.nih.gov/38078573/)]
12. Zou Q, Qu K, Luo Y, Yin D, Ju Y, Tang H. Predicting diabetes mellitus with machine learning techniques. *Front Genet* 2018;9:515 [FREE Full text] [doi: [10.3389/fgene.2018.00515](https://doi.org/10.3389/fgene.2018.00515)] [Medline: [30459809](https://pubmed.ncbi.nlm.nih.gov/30459809/)]
13. Dagliati A, Marini S, Sacchi L, Cogni G, Teliti M, Tibollo V, et al. Machine learning methods to predict diabetes complications. *J Diabetes Sci Technol* 2018;12(2):295-302 [FREE Full text] [doi: [10.1177/1932296817706375](https://doi.org/10.1177/1932296817706375)] [Medline: [28494618](https://pubmed.ncbi.nlm.nih.gov/28494618/)]
14. Islam MS, Qaraqe MK, Belhaouari S, Petrovski G. Long term HbA1c prediction using multi-stage CGM data analysis. *IEEE Sens J* 2021;21(13):15237-15247. [doi: [10.1109/JSEN.2021.3073974](https://doi.org/10.1109/JSEN.2021.3073974)]
15. Vincent P, Larochelle H, Bengio Y, Manzagol P. Extracting and composing robust features with denoising autoencoders. : Association for Computing Machinery; 2008 Presented at: Proceedings of the 25th international conference on Machine learning (ICML '08); 2008; Canada p. 1096-1103.

16. Jawed S, Grabocka J, Schmidt-Thieme L. Self-supervised learning for semi-supervised time series classification. : Springer; 2020 Presented at: Advances in Knowledge Discovery and Data Mining (PAKDD 2020); May 11, 2020; Asia p. 12084. [doi: [10.1007/978-3-030-47426-3_39](https://doi.org/10.1007/978-3-030-47426-3_39)]
17. Tipirneni S, Reddy CK. Self-supervised transformer for sparse and irregularly sampled multivariate clinical time-series. *ACM Trans Knowl Discov Data* 2022;16(6):1-17. [doi: [10.1145/3516367](https://doi.org/10.1145/3516367)]
18. Chang BY, Naiel MA, Wardell S, Kleinikink S, Zelek JS. Time-series causality with missing data. *J Comp Vis Imag Sys* 2021;6(1):1-4. [doi: [10.15353/jcvis.v6i1.3552](https://doi.org/10.15353/jcvis.v6i1.3552)]
19. Theissler A, Spinnato F, Schlegel U, Guidotti R. Explainable AI for time series classification: a review, taxonomy and research directions. : IEEE; 2022 Presented at: IEEE Access; September 19, 2022; Australia p. 100700-100724. [doi: [10.1109/access.2022.3207765](https://doi.org/10.1109/access.2022.3207765)]
20. Maassen O, Fritsch S, Palm J, Deffge S, Kunze J, Marx G, et al. Future medical artificial intelligence application requirements and expectations of physicians in German university hospitals: web-based survey. *J Med Internet Res* 2021;23(3):e26646 [FREE Full text] [doi: [10.2196/26646](https://doi.org/10.2196/26646)] [Medline: [33666563](https://pubmed.ncbi.nlm.nih.gov/33666563/)]
21. Diprose WK, Buist N, Hua N, Thurier Q, Shand G, Robinson R. Physician understanding, explainability, and trust in a hypothetical machine learning risk calculator. *J Am Med Inform Assoc* 2020;27(4):592-600 [FREE Full text] [doi: [10.1093/jamia/ocz229](https://doi.org/10.1093/jamia/ocz229)] [Medline: [32106285](https://pubmed.ncbi.nlm.nih.gov/32106285/)]
22. Kodera S, Ninomiya K, Sawano S, Katsushika S, Shinohara H, Akazawa H, et al. Patient awareness survey on medical AI. 2022 Presented at: The 36th Annual Conference of the Japanese Society for Artificial Intelligence; June 14, 2022; Kyoto, Japan.
23. Ossa LA, Starke G, Lorenzini G, Vogt JE, Shaw DM, Elger BS. Re-focusing explainability in medicine. *Digit Health* 2022;8:20552076221074488 [FREE Full text] [doi: [10.1177/20552076221074488](https://doi.org/10.1177/20552076221074488)] [Medline: [35173981](https://pubmed.ncbi.nlm.nih.gov/35173981/)]
24. Allam A, Feuerriegel S, Rebhan M, Krauthammer M. Analyzing patient trajectories with artificial intelligence. *J Med Internet Res* 2021;23(12):e29812 [FREE Full text] [doi: [10.2196/29812](https://doi.org/10.2196/29812)] [Medline: [34870606](https://pubmed.ncbi.nlm.nih.gov/34870606/)]
25. Richens JG, Lee CM, Johri S. Improving the accuracy of medical diagnosis with causal machine learning. *Nat Commun* 2020;11(1):3923 [FREE Full text] [doi: [10.1038/s41467-020-17419-7](https://doi.org/10.1038/s41467-020-17419-7)] [Medline: [32782264](https://pubmed.ncbi.nlm.nih.gov/32782264/)]
26. Rojat T, Puget R, Filliat D, Del SJ, Gelin R, Díaz-Rodríguez N. Explainable artificial intelligence (XAI) on TimeSeries data: a survey. *arXiv Preprint* posted online on April 2, 2021. [doi: [10.48550/arXiv.2104.00950](https://doi.org/10.48550/arXiv.2104.00950)]
27. Bommasani R, Hudson D, Adeli E. On the opportunities and risks of foundation models. *arXiv Preprint* posted online on August 16, 2021. [doi: [10.48550/arXiv.2108.07258](https://doi.org/10.48550/arXiv.2108.07258)]
28. Lin T, Wang Y, Liu X, Qiu X. A survey of transformers. *AI Open* 2022;3:111-132. [doi: [10.1016/j.aiopen.2022.10.001](https://doi.org/10.1016/j.aiopen.2022.10.001)]
29. Lakew S, Cettolo M, Federico M. A comparison of transformer and recurrent neural networks on multilingual neural machine translation. 2018 Presented at: Proceedings of the 27th International Conference on Computational Linguistics; 2018; Barcelona, Spain (Online) p. 641-652. [doi: [10.18653/v1/2020.coling-tutorials.3](https://doi.org/10.18653/v1/2020.coling-tutorials.3)]
30. Lu K, Xu Y, Yang Y. Comparison of the potential between transformer and CNN in image classification. 2021 Presented at: ICMLCA 2021; 2nd International Conference on Machine Learning and Computer Application; December 17, 2021; Shenyang, China p. 1-6.
31. Zerveas G, Jayaraman S, Patel D, Bhamidipaty A, Eickhoff C. A transformer-based framework for multivariate time series representation learning. 2021 Presented at: Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining (KDD '21); August 14, 2021; New York p. 2114-2124. [doi: [10.1145/3447548.3467401](https://doi.org/10.1145/3447548.3467401)]
32. Zhou H, Zhang S, Peng J, Zhang S, Li J, Xiong H, et al. Informer: beyond efficient transformer for long sequence time-series forecasting. 2021 Presented at: Proceedings of the AAAI Conference on Artificial Intelligence; February 20, 2024; USA p. 11106-11115. [doi: [10.1609/aaai.v35i12.17325](https://doi.org/10.1609/aaai.v35i12.17325)]
33. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A, et al. Attention is all you need. *Adv Neural Inf Process Syst* 2017;30. [doi: [10.48550/arXiv.1706.03762](https://doi.org/10.48550/arXiv.1706.03762)]
34. Vig J, Belinkov Y. Analyzing the structure of attention in a transformer language model. 2019 Presented at: Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP; 2019; Florence, Italy p. 63-76. [doi: [10.18653/v1/w19-4808](https://doi.org/10.18653/v1/w19-4808)]
35. Voita E, Talbot D, Moiseev F, Sennrich R, Titov I. Analyzing multi-head self-attention: specialized heads do the heavy lifting, the rest can be pruned. 2019 Presented at: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics; 2019; Florence, Italy p. 5797-5808. [doi: [10.18653/v1/p19-1580](https://doi.org/10.18653/v1/p19-1580)]
36. Chaddad A, Peng J, Xu J, Bouridane A. Survey of explainable AI techniques in healthcare. *Sensors (Basel)* 2023;23(2):634 [FREE Full text] [doi: [10.3390/s23020634](https://doi.org/10.3390/s23020634)] [Medline: [36679430](https://pubmed.ncbi.nlm.nih.gov/36679430/)]
37. Hosmer D, Lemeshow S. *Applied Logistic Regression*. New York: John Wiley and Sons; 2000:160-164.
38. Lv X, Luo J, Huang W, Guo H, Bai X, Yan P, et al. Identifying diagnostic indicators for type 2 diabetes mellitus from physical examination using interpretable machine learning approach. *Front Endocrinol (Lausanne)* 2024;15:1376220 [FREE Full text] [doi: [10.3389/fendo.2024.1376220](https://doi.org/10.3389/fendo.2024.1376220)] [Medline: [38562414](https://pubmed.ncbi.nlm.nih.gov/38562414/)]
39. Uchitachimoto G, Sukegawa N, Kojima M, Kagawa R, Oyama T, Okada Y, et al. Data collaboration analysis in predicting diabetes from a small amount of health checkup data. *Sci Rep* 2023;13(1):11820 [FREE Full text] [doi: [10.1038/s41598-023-38932-x](https://doi.org/10.1038/s41598-023-38932-x)] [Medline: [37479701](https://pubmed.ncbi.nlm.nih.gov/37479701/)]

40. Choi SG, Oh M, Park D, Lee B, Lee Y, Jee SH, et al. Comparisons of the prediction models for undiagnosed diabetes between machine learning versus traditional statistical methods. *Sci Rep* 2023;13(1):13101 [FREE Full text] [doi: [10.1038/s41598-023-40170-0](https://doi.org/10.1038/s41598-023-40170-0)] [Medline: [37567907](https://pubmed.ncbi.nlm.nih.gov/37567907/)]
41. Agliata A, Giordano D, Bardozzo F, Bottiglieri S, Facchiano A, Tagliaferri R. Machine learning as a support for the diagnosis of type 2 diabetes. *Int J Mol Sci* 2023;24(7):6775 [FREE Full text] [doi: [10.3390/ijms24076775](https://doi.org/10.3390/ijms24076775)] [Medline: [37047748](https://pubmed.ncbi.nlm.nih.gov/37047748/)]
42. Dutta A, Hasan MK, Ahmad M, Awal MA, Islam MA, Masud M, et al. Early prediction of diabetes using an ensemble of machine learning models. *Int J Environ Res Public Health* 2022;19(19):12378 [FREE Full text] [doi: [10.3390/ijerph191912378](https://doi.org/10.3390/ijerph191912378)] [Medline: [36231678](https://pubmed.ncbi.nlm.nih.gov/36231678/)]
43. Waki K, Fujita H, Uchimura Y, Omae K, Aramaki E, Kato S, et al. DialBetics: a novel smartphone-based self-management support system for type 2 diabetes patients. *J Diabetes Sci Technol* 2014;8(2):209-215 [FREE Full text] [doi: [10.1177/1932296814526495](https://doi.org/10.1177/1932296814526495)] [Medline: [24876569](https://pubmed.ncbi.nlm.nih.gov/24876569/)]
44. Kurasawa H, Hayashi K, Fujino A, Takasugi K, Haga T, Waki K, et al. Machine-learning-based prediction of a missed scheduled clinical appointment by patients with diabetes. *J Diabetes Sci Technol* 2016;10(3):730-736 [FREE Full text] [doi: [10.1177/1932296815614866](https://doi.org/10.1177/1932296815614866)] [Medline: [26555782](https://pubmed.ncbi.nlm.nih.gov/26555782/)]
45. Kurasawa H, Waki K, Chiba A, Seki T, Hayashi K, Fujino A, et al. Treatment discontinuation prediction in patients with diabetes using a ranking model: machine learning model development. *JMIR Bioinform Biotech* 2022;3(1):e37951. [doi: [10.2196/37951](https://doi.org/10.2196/37951)]
46. Kazijevs M, Samad MD. Deep imputation of missing values in time series health data: a review with benchmarking. *J Biomed Inform* 2023;144:104440. [doi: [10.1016/j.jbi.2023.104440](https://doi.org/10.1016/j.jbi.2023.104440)] [Medline: [37429511](https://pubmed.ncbi.nlm.nih.gov/37429511/)]
47. Tashman LJ. Out-of-sample tests of forecasting accuracy: an analysis and review. *Int J Forecast* 2000;16(4):437-450. [doi: [10.1016/s0169-2070\(00\)00065-0](https://doi.org/10.1016/s0169-2070(00)00065-0)]
48. Seto H, Oyama A, Kitora S, Toki H, Yamamoto R, Kotoku J, et al. Author correction: gradient boosting decision tree becomes more reliable than logistic regression in predicting probability for diabetes with big data. *Sci Rep* 2022;12(1):22599 [FREE Full text] [doi: [10.1038/s41598-022-27052-7](https://doi.org/10.1038/s41598-022-27052-7)] [Medline: [36585468](https://pubmed.ncbi.nlm.nih.gov/36585468/)]
49. Rufo DD, Debelee TG, Ibenthal A, Negera WG. Diagnosis of diabetes mellitus using gradient boosting machine (LightGBM). *Diagnostics (Basel)* 2021;11(9):1714 [FREE Full text] [doi: [10.3390/diagnostics11091714](https://doi.org/10.3390/diagnostics11091714)] [Medline: [34574055](https://pubmed.ncbi.nlm.nih.gov/34574055/)]
50. Information about the current study. URL: https://www.h.u-tokyo.ac.jp/patient/depts/taisha/pdf/pa_md_md_info-04.pdf [accessed 2023-01-03]
51. Henighan T, Kaplan J, Katz M. Scaling laws for autoregressive generative modeling. arXiv Preprint posted online on October 28, 2020. [doi: [10.48550/arXiv.2001.08361](https://doi.org/10.48550/arXiv.2001.08361)]
52. Yang Z, Mitra A, Liu W, Berlowitz D, Yu H. TransformEHR: transformer-based encoder-decoder generative model to enhance prediction of disease outcomes using electronic health records. *Nat Commun* 2023;14(1):7857 [FREE Full text] [doi: [10.1038/s41467-023-43715-z](https://doi.org/10.1038/s41467-023-43715-z)] [Medline: [38030638](https://pubmed.ncbi.nlm.nih.gov/38030638/)]
53. Li Y, Rao S, Solares JRA, Hassaine A, Ramakrishnan R, Canoy D, et al. BEHRT: transformer for electronic health records. *Sci Rep* 2020;10(1):7155 [FREE Full text] [doi: [10.1038/s41598-020-62922-y](https://doi.org/10.1038/s41598-020-62922-y)] [Medline: [32346050](https://pubmed.ncbi.nlm.nih.gov/32346050/)]
54. Kraljevic Z, Bean D, Shek A, Bendayan R, Hemingway H, Yeung JA, et al. Foresight-a generative pretrained transformer for modelling of patient timelines using electronic health records: a retrospective modelling study. *Lancet Digit Health* 2024;6(4):e281-e290 [FREE Full text] [doi: [10.1016/S2589-7500\(24\)00025-6](https://doi.org/10.1016/S2589-7500(24)00025-6)] [Medline: [38519155](https://pubmed.ncbi.nlm.nih.gov/38519155/)]
55. Feuerriegel S, Frauen D, Melnychuk V, Schweisthal J, Hess K, Curth A, et al. Causal machine learning for predicting treatment outcomes. *Nat Med* 2024;30(4):958-968. [doi: [10.1038/s41591-024-02902-1](https://doi.org/10.1038/s41591-024-02902-1)] [Medline: [38641741](https://pubmed.ncbi.nlm.nih.gov/38641741/)]

Abbreviations

- AUC:** area under the curve
- EHR:** electronic health record
- HbA_{1c}:** hemoglobin A_{1c}
- ML:** machine learning
- PR:** precision-recall
- ROC:** receiver operating characteristic
- T2D:** type 2 diabetes
- XAI:** explainable artificial intelligence

Edited by K El Emam, B Malin; submitted 24.01.24; peer-reviewed by G Lim, A Jafarizadeh; comments to author 21.03.24; revised version received 21.04.24; accepted 31.05.24; published 18.07.24.

Please cite as:

*Kurasawa H, Waki K, Seki T, Chiba A, Fujino A, Hayashi K, Nakahara E, Haga T, Noguchi T, Ohe K
Enhancing Type 2 Diabetes Treatment Decisions With Interpretable Machine Learning Models for Predicting Hemoglobin A1c
Changes: Machine Learning Model Development*

JMIR AI 2024;3:e56700

URL: <https://ai.jmir.org/2024/1/e56700>

doi: [10.2196/56700](https://doi.org/10.2196/56700)

PMID: [39024008](https://pubmed.ncbi.nlm.nih.gov/39024008/)

©Hisashi Kurasawa, Kayo Waki, Tomohisa Seki, Akihiro Chiba, Akinori Fujino, Katsuyoshi Hayashi, Eri Nakahara, Tsuneyuki Haga, Takashi Noguchi, Kazuhiko Ohe. Originally published in JMIR AI (<https://ai.jmir.org>), 18.07.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR AI, is properly cited. The complete bibliographic information, a link to the original publication on <https://www.ai.jmir.org/>, as well as this copyright and license information must be included.

Original Paper

Identifying Frailty in Older Adults Receiving Home Care Assessment Using Machine Learning: Longitudinal Observational Study on the Role of Classifier, Feature Selection, and Sample Size

Cheng Pan¹, MPhil; Hao Luo², PhD; Gary Cheung³, PhD; Huiquan Zhou², PhD; Reynold Cheng¹, PhD; Sarah Cullum³, PhD; Chuan Wu¹, PhD

¹Department of Computer Science, The University of Hong Kong, Hong Kong, China (Hong Kong)

²Department of Social Work and Social Administration, The University of Hong Kong, Hong Kong, China (Hong Kong)

³Department of Psychological Medicine, School of Medicine, The University of Auckland, Auckland, New Zealand

Corresponding Author:

Hao Luo, PhD

Department of Social Work and Social Administration

The University of Hong Kong

CJT 521, Jockey Club Tower

Pokfulam Road

Hong Kong

China (Hong Kong)

Phone: 852 68421252

Email: haoluo@hku.hk

Abstract

Background: Machine learning techniques are starting to be used in various health care data sets to identify frail persons who may benefit from interventions. However, evidence about the performance of machine learning techniques compared to conventional regression is mixed. It is also unclear what methodological and database factors are associated with performance.

Objective: This study aimed to compare the mortality prediction accuracy of various machine learning classifiers for identifying frail older adults in different scenarios.

Methods: We used deidentified data collected from older adults (65 years of age and older) assessed with interRAI-Home Care instrument in New Zealand between January 1, 2012, and December 31, 2016. A total of 138 interRAI assessment items were used to predict 6-month and 12-month mortality, using 3 machine learning classifiers (random forest [RF], extreme gradient boosting [XGBoost], and multilayer perceptron [MLP]) and regularized logistic regression. We conducted a simulation study comparing the performance of machine learning models with logistic regression and interRAI Home Care Frailty Scale and examined the effects of sample sizes, the number of features, and train-test split ratios.

Results: A total of 95,042 older adults (median age 82.66 years, IQR 77.92-88.76; n=37,462, 39.42% male) receiving home care were analyzed. The average area under the curve (AUC) and sensitivities of 6-month mortality prediction showed that machine learning classifiers did not outperform regularized logistic regressions. In terms of AUC, regularized logistic regression had better performance than XGBoost, MLP, and RF when the number of features was ≤ 80 and the sample size $\leq 16,000$; MLP outperformed regularized logistic regression in terms of sensitivities when the number of features was ≥ 40 and the sample size ≥ 4000 . Conversely, RF and XGBoost demonstrated higher specificities than regularized logistic regression in all scenarios.

Conclusions: The study revealed that machine learning models exhibited significant variation in prediction performance when evaluated using different metrics. Regularized logistic regression was an effective model for identifying frail older adults receiving home care, as indicated by the AUC, particularly when the number of features and sample sizes were not excessively large. Conversely, MLP displayed superior sensitivity, while RF exhibited superior specificity when the number of features and sample sizes were large.

(JMIR AI 2024;3:e44185) doi:[10.2196/44185](https://doi.org/10.2196/44185)

KEYWORDS

machine learning; logistic regression; frailty; older adults; home care; sample size; features; data set; model; home care; mortality prediction; assessment

Introduction

Frailty is a syndrome characterized by an increased vulnerability to adverse health outcomes, including falling, hospitalization, physical decline, and mortality [1]. Frailty should be detected as early as possible since it is potentially preventable and treatable [2]. In community settings, timely identification of frailty allows the implementation of early interventions that could reduce care costs and improve the “ability of older persons to age in place” [3]. In clinical and long-term care settings, identifying frail older adults could facilitate more individualized and tailored health care planning [4,5]. Therefore, efficient and accurate clinical tools are pivotal to the early identification of frailty among at-risk older adults.

Numerous methods have been applied to measure frailty. A recent systematic review identified 21 conceptual definitions and 59 operational definitions of frailty from 68 studies [6]. This review concluded that definitions of frailty can be classified into 3 categories focusing on different dimensions. The first is represented by the Cardiovascular Health Study (CHS) Index based on Fried’s “frailty phenotype” model, which focuses on the physical dimensions of frailty [7-10]. The second category is represented by the Frailty Index, originally proposed by Rockwood and Mitnitski [11,12], which considers frailty as a syndrome capturing the accumulative gradient of deficits. This category of definitions covers other dimensions of frailty, including cognitive, psychological, nutritional, and social factors [11,13]. The third category considers the social dimension of frailty, which has a significant relationship with undesirable adverse health outcomes [14-16]. Despite differences in theoretical frameworks adopted by different frailty measures, existing frailty indices are typically constructed by summing up the number of deficits or scores of assessment items using equal weighting. Arguably, different deficits from various domains may impact overall frailty status differently, and these differences should be considered when measuring frailty. In addition to accounting for the multifactorial nature of frailty, a successful definition of frailty [12] must demonstrate satisfactory criterion validity. Since frailty is noncontroversially linked with vulnerability, a valid measure of frailty must accurately predict adverse outcomes, such as death, institutionalization, hospitalization, physical decline, and falls. Mortality is the most objective measure that is less susceptible to measurement error and, thus, is the most widely used outcome for assessing the predictive validity of frailty measures [9,17-20].

Routinely collected data from health information systems have become increasingly available in recent years, and clinical big data analytics featured by machine learning techniques are ever-evolving [21-23]. In contrast to conventional regression approaches, classifiers used in machine learning, such as random forest (RF), support vector machines, and neural networks, have the advantages of learning and generating predictions by examining large-scale databases of complex clinical information [18,20,24-26]. Therefore, it is reasonable to hypothesize that

applying machine learning techniques to large-scale data collected from health information systems can improve the accuracy of mortality prediction for identifying frail older persons who may benefit from early interventions. However, the literature remains unclear whether machine learning techniques can outperform conventional regression models in identifying frail older adults [18,19,27].

In this study, we used routinely collected health information of people receiving home care in New Zealand from interRAI-Home Care (interRAI-HC) assessment to examine the performance of various machine learning classifiers in mortality prediction for identifying frailty. In this study, we conducted a simulation study to address the following research questions: (1) does the performance of machine learning models exceed that of the interRAI-HC Frailty Scale, which was developed using conventional regression models [28], in identifying frailty? (2) what are the performances of different machine learning models? and (3) what are the effects of sample size, number of features, and the ratio of training to test data on predictive accuracy?

Methods

Data Source and Participants

In this retrospective observational study, we used deidentified health information routinely collected from older adults assessed using the interRAI-HC assessment (version 9.1). The interRAI-HC assessment was developed by a network of health researchers in over 35 countries [29]. interRAI assessments are mandatory in aged residential care and home and community services for older people living in the community in New Zealand. Our participants were from all 20 District Health Boards in New Zealand and included all community-dwelling older adults who were receiving public-funded home care or assessed for long-term aged residential care. Trained interRAI assessors collect comprehensive health information on older adults, including their demographic, clinical, psychosocial, and functional details. The interRAI-HC assessment embeds over 100 potential deficits of older adults that can be used to identify frailty. Table S1 in [Multimedia Appendix 1](#) summarizes the variables used for identifying frail older adults. Ethnicity was not included to increase generalizability beyond New Zealand.

We included adults 65 years of age or older for whom at least 1 interRAI-HC assessment had been completed between January 1, 2012, and December 31, 2016. Only the most recent interRAI-HC assessment (defined as the index assessment) of each individual within this period was used in the analysis and the date of the most recent assessment was defined as the index date. The individuals were followed from the index date until the date of death or December 31, 2019, whichever came first.

Ethical Considerations

The University of Auckland Human Participant Ethics Committee provided ethics approval for this study (023801).

Measures

Outcomes

Outcomes of interest were 6-month and 12-month mortality. Mortality data were retrieved from the Ministry of Health Mortality Dataset that contains information of all registered deaths in New Zealand. These two-time points were chosen because (1) older adults receiving home care are associated with a higher risk of mortality and shorter survival compared with their counterparts who are not receiving home care and (2) these are outcomes commonly used in previous studies examining the association between frailty and mortality [30-33] and few previous studies using interRAI data [34-36].

Features Used in Machine Learning Models

Features of interest included 138 interRAI-HC assessment items covering 11 broad domains, demographics, cognition, communication and vision, mood and behavior, psychosocial well-being, functional status, continence, disease diagnoses, health conditions, oral and nutrition status, and skin conditions. Table S1 in [Multimedia Appendix 1](#) presents the details of features used to identify frail older individuals.

Assessment items that had a missing percentage of over 10% were excluded from this study. Multiple interRAI-HC assessment variables with a response indicating that the activity did not occur during the assessment were considered missing, and the missing data imputation was implemented for these responses.

Established Frailty Scales (Benchmark)

The interRAI-HC Frailty Scale was used as the benchmark for evaluating the predictive performance of machine learning algorithms. The interRAI-HC Frailty Scale was developed and validated using assessments collected from multiple and diverse countries worldwide [28]. Table S2 in [Multimedia Appendix 1](#) summarizes the variables used in constructing the interRAI-HC Frailty Scale.

Machine Learning and Logistic Regression Models

We applied 3 state-of-the-art machine learning models and regularized logistic regression to predict 6-month and 12-month mortality using the features available from interRAI-HC. The RF is a machine learning algorithm that uses decision trees [37]. The RF provides highly accurate predictions with a very large number of input variables [38]. The eXtreme Gradient Boosting

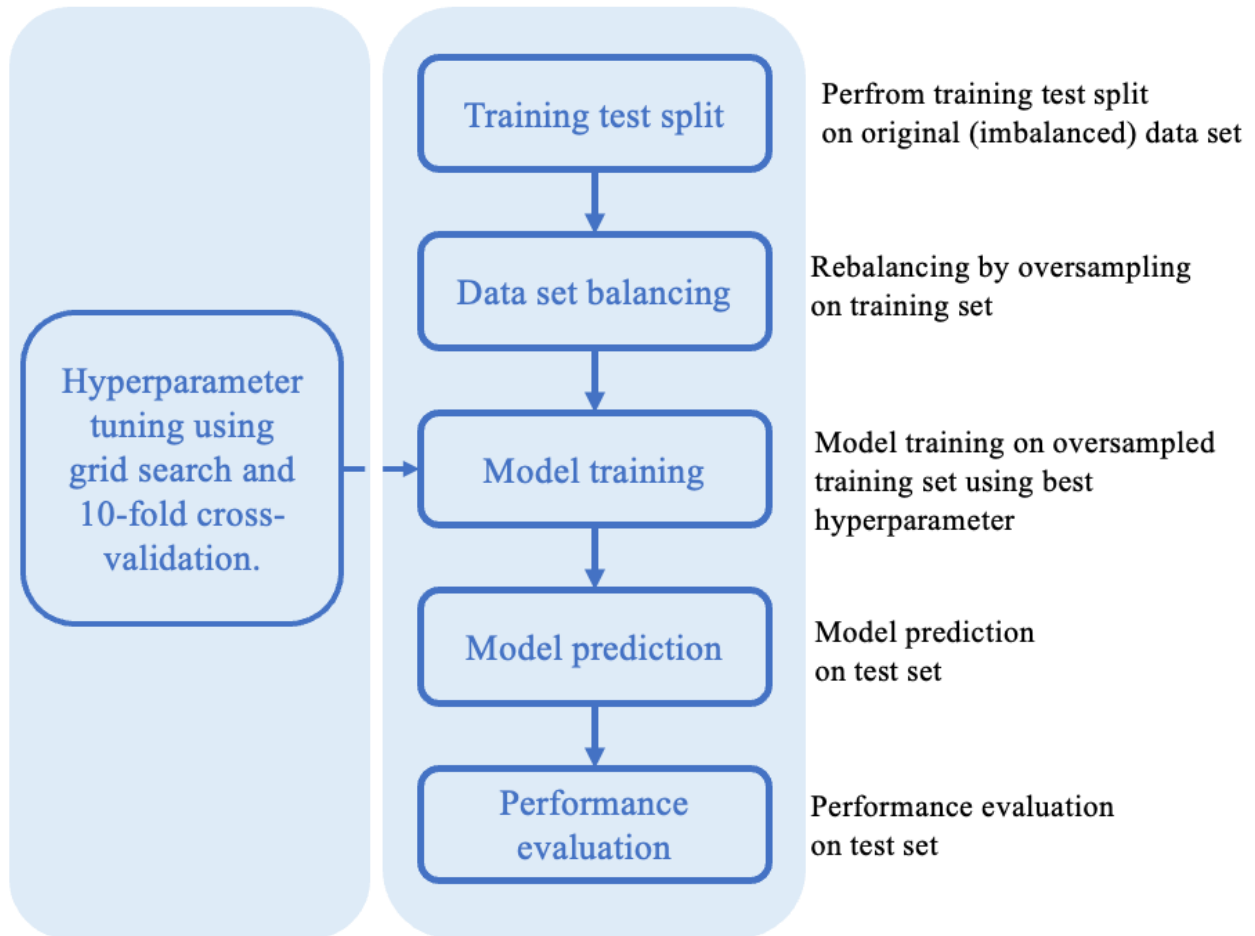
(XGBoost) is an optimized algorithm designed to implement parallel tree boosting that can predict results extremely efficiently and accurately based on its scalability and efficiency in all scenarios [39]. Multilayer perceptron (MLP) is one of the most popular paradigms of artificial neural networks. MLP decreases the output error by adjusting the weights of predictive variables through an iterative learning process [40].

Regularized logistic regression is a variant of logistic regression using regularization to prevent overfitting and improve the performance of logistic regression. Two popular types of regularized logistic regressions are Least Absolute Shrinkage and Selection Operator (LASSO) regularization with the L1 penalty [41] and Ridge regularization with the L2 penalty [42].

In this study, we implemented hyperparameter tuning to regularize logistic regression (hereafter referred to as logistic regression), RF, MLP, and XGBoost by performing a randomized grid search using all home care (HC) assessment items. The best hyperparameters for each classifier were determined by 10-fold cross-validation (Table S5 in [Multimedia Appendix 1](#)). We used iterative imputation [43] to handle the missing values and the default threshold of 0.5 was used in training [27]. We conducted a sensitivity analysis to compare the performance of the models with and without imputation in selected conditions, that is, only the minimum and maximum sample sizes and the number of features were selected for comparison due to the expensive computation power required.

The preliminary results suggested that our data are imbalanced, as the majority of individuals survived within 6 or 12 months. We therefore rebalanced the training data (but not the test data) using random oversampling [44], while keeping the test data unchanged. Our primary findings are presented with the results obtained after rebalancing the data. The results using the original imbalanced data set can be found in [Multimedia Appendix 1](#). Specifically, to initiate the hyperparameter tuning process, we performed hyperparameter tuning using grid search. For each combination of hyperparameters, within each iteration of the 10-fold cross-validation loop, we applied oversampling to the training set, and the model was trained on the oversampled training set using the current combination of hyperparameters. The model's performance was evaluated on the validation set. After all combinations of hyperparameters have been evaluated, we selected the combination that gave the best average performance. The process of data preprocessing, training, prediction, and evaluation is illustrated in [Figure 1](#).

Figure 1. Illustration of the process of data preprocessing, training, prediction, and evaluation.



Simulation Design

We conducted a Monte Carlo simulation to compare the performance of different machine learning methods and logistic regression under different experimental conditions, characterized by different sample sizes, the number of features, and training test split ratios. There were 72 experimental conditions for each model (4 sample sizes, 6 feature numbers, and 3 training test split ratios). Each of these conditions was repeated 1000 times to assess their variability. We used sample sizes equaling 1000, 4000, 16,000, and 95,042; the number of features equaling 10, 20, 30, 40, 80, and 138; and training test split ratios equaling 7:3, 8:2, and 9:1 in our simulation. We selected these sample sizes and feature numbers because they are commonly encountered in existing studies on frailty measurement [17,19,45-48] and are values that are testable using the current database. The training split ratios are widely used in studies using machine learning [18,27,36,49,50]. We chose a limited number under each domain to keep the simulations to a manageable scale.

Evaluation of Model Performance

We randomly split the data into a training sample and a test sample with different training test ratios. We evaluated model performances using the test sample. The discrimination ability of each classifier was measured by the area under the curve (AUC) [51], sensitivity, (also referred to as the true positive rate), and specificity (also known as the true negative rate) as

the primary criteria because these are criteria widely accepted by the clinicians. Since frailty is reversible and may be attenuated by noninvasive interventions such as exercise, reduction of polypharmacy, and adequate nutrition [52], high sensitivity is viewed as more important than high specificity in this context if a trade-off needs to be made. F_1 -score [53], accuracy and precision (also called positive predictive value) [47,54,55] were also constructed and assessed to allow comparisons with studies that reported only these outcomes. Note, that as each experimental condition was repeated 1000 times to address the potential impact of randomization, we computed the mean and SDs of all performance indices across 1000 replications. The 95% CI for the performance metrics was computed from 1000 runs for each scenario.

Results

We included 95,042 older adults after excluding 4676 individuals who were younger than 65 years of age and 51 individuals with incorrect records (eg, the date of death was earlier than the assessment date, invalid date of birth, or an incorrect assessment date). Table 1 summarizes the characteristics of study subjects, stratified by whether the person died within 6 months. About half of the subjects were aged between 80 and 89 years (80-84 years: $n=21,947$, 23.09%; 85-89 years: $n=23,906$, 25.15%). Women accounted for 57,580 (60.58%) of the sample, and 83,590 (87.95%) were European.

A total of 12,401 (13.05%) subjects died within 6 months following the index assessment. Table S19 in [Multimedia Appendix 1](#) documents the characteristics of the study subjects, stratified by whether the person died within 1 year.

Table S4 in [Multimedia Appendix 1](#) presents the results of the sensitivity analysis comparing the performance of the models with and without imputation. The findings suggest that the data imputation was necessary as the imputed data set outperformed the unimputed data set in most of the conditions tested.

After comparing the performance of penalty terms none, L1, and L2, the LASSO regression regularization (L1) and Ridge regularization (L2) were used in 6-month and 12-month mortality prediction, respectively. We compared the average AUC of each classifier as the number of features increased for 6-month mortality prediction ([Figure 2](#)). Overall, the performance of all methods improved considerably as the number of features increased. Specifically, in most scenarios, when the number of features increased to 30, four classifiers demonstrated significantly higher AUC than the interRAI-HC Frailty Scale. LASSO regression generally demonstrated higher or comparable AUC than RF, MLP, and XGBoost. However, in the specific scenario where the sample size was 95,042 and the number of features was 40 or less, MLP showed a slightly better average AUC than LASSO regression. In addition, when the sample size was 95,042, and the number of features increased to 138, XGBoost achieved the highest average AUC of 0.79 (95% CI 0.79-0.80).

[Figure 3](#) shows the average sensitivities across all experimental conditions. The 3 machine learning classifiers and LASSO regression had lower sensitivities than the interRAI-HC frailty scale when the sample size was 1000. As the sample size increased to 4000 and the number of features increased to 20, MLP and LASSO regression outperformed the benchmark scale with the highest average sensitivity of 0.77 (95% CI 0.72-0.79) observed in MLP when the sample size was 95,042, and the number of features was 138. Meanwhile, all classifiers demonstrated higher average specificities than the interRAI-HC

Frailty Scale in all scenarios ([Figure 4](#)). The RF and XGBoost demonstrated higher specificities than LASSO regression, with RF achieving the highest average specificities of 0.98 (95% CI 0.98-0.98) when the sample size was 95,042 and the number of features was 138.

Based on the simulation results, it was observed that the test size ratios did not have a significant impact on the average AUC, sensitivities, and specificities, as shown in [Figure 5](#). The 12-month and 6-month mortality predictions were comparable ([Figures S1-S4 in Multimedia Appendix 1](#)). However, the overall performance of logistic regression on the 12-month mortality prediction was worse than the 6-month prediction. Compared to the 6-month mortality prediction, machine learning classifiers performed slightly better average sensitivities and worse average AUCs and specificities on 12-month mortality prediction. [Tables S5-S18 and S20-S33 in Multimedia Appendix 1](#) summarize AUC, sensitivity, specificity, F_1 -score, accuracy, and precision.

Our simulation was also conducted on the imbalanced data set, and we observed a similar result in terms of average AUCs. Regularized logistic regression had a higher AUC than XGBoost, MLP, and RF, especially when the number of features was less than or equal to 80 and the sample size was less than or equal to 16,000. However, as the number of features and sample sizes increased, XGBoost slightly outperformed regularized logistic regression. In terms of sensitivities, regularized logistic regression significantly outperformed machine learning classifiers in all scenarios, while machine learning classifiers had higher specificities than regularized logistic regression in all scenarios. Additionally, the findings for 12-month and 6-month mortality prediction were similar. However, machine learning classifiers performed slightly better in average sensitivities, but worse in average AUCs and specificities for 12-month mortality prediction compared to 6-month mortality prediction. [Multimedia Appendix 1](#) has been included to summarize the results of the imbalanced data set ([Tables S34-S62 and Figures S9-S12 in Multimedia Appendix 1](#)).

Table 1. Sample characteristics of 6-month mortality.

Characteristics	HC ^a (N=95,042)	6-month deceased (n=12,401)	6-month survived (n=82,641)
Age (years)			
65-69, n (%)	5906 (6.21)	693 (5.59)	5213 (6.31)
70-74, n (%)	9623 (10.12)	1065 (8.59)	8558 (10.36)
75-79, n (%)	15,284 (16.08)	1770 (14.27)	13,514 (16.35)
80-84, n (%)	21,947 (23.09)	2662 (21.47)	19,285 (23.34)
85-89, n (%)	23,906 (25.15)	3312 (26.71)	20,594 (24.92)
90-94, n (%)	14,370 (15.12)	2160 (17.42)	12,210 (14.77)
95-99, n (%)	3594 (3.78)	654 (5.27)	2940 (3.56)
≥100, n (%)	412 (0.43)	85 (0.69)	327 (0.40)
Mean (SD)	82.66 (7.61)	83.59 (7.71)	82.52 (7.59)
Gender, n (%)			
Female	57,580 (60.58)	6362 (51.30)	51,218 (61.98)
Male	37,462 (39.42)	6039 (48.70)	31,423 (38.02)
Ethnicity, n (%)			
European	83,590 (87.95)	11,128 (89.73)	72,462 (87.68)
Maori	5321 (5.60)	730 (5.89)	4591 (5.56)
Pacific Island	2948 (3.10)	267 (2.15)	2681 (3.24)
Asian	2304 (2.42)	197 (1.59)	2107 (2.55)
Middle eastern or Latin American or African	352 (0.37)	25 (0.20)	327 (0.40)
Other ethnicity	527 (0.55)	54 (0.44)	473 (0.57)
Marital status, n (%)			
Married or civil union or de facto	82,401 (86.70)	10,936 (88.19)	71,465 (86.48)
Never married	4486 (4.72)	539 (4.35)	3947 (4.78)
Widowed	2116 (2.23)	240 (1.94)	1876 (2.27)
Separated or divorced	5999 (6.31)	683 (5.51)	5316 (6.43)
Others	40 (0.04)	3 (0.02)	37 (0.04)

^aHC: home care.

Figure 2. Average AUCs of classifiers and frailty scale for 6-month mortality prediction on balanced data set. AUC: area under the curve; HC: home care; LR: logistic regression; MLP: multilayer perceptron; XGBoost: extreme gradient boosting.

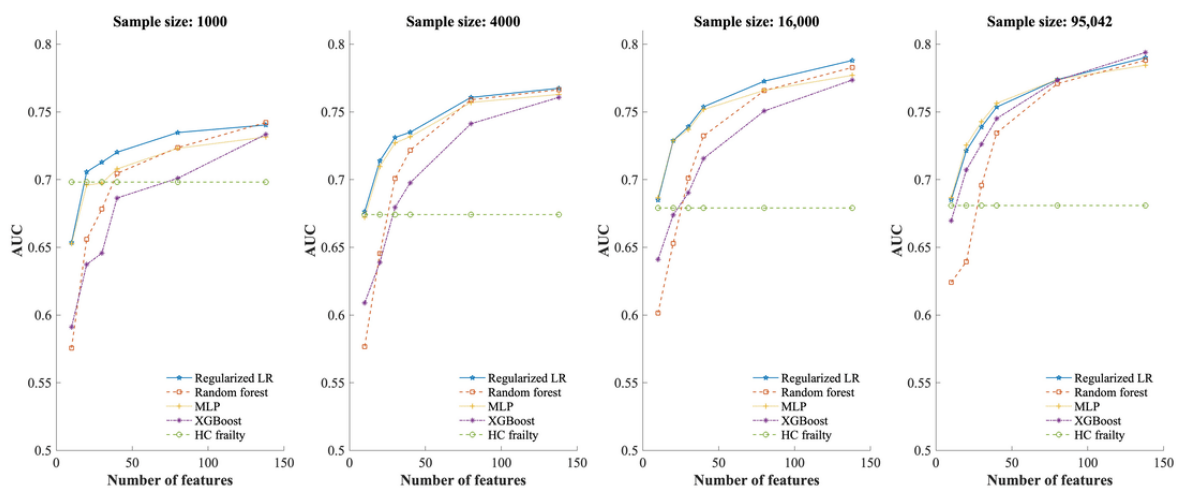


Figure 3. Average sensitivities of classifiers and frailty scale for 6-month mortality prediction on balanced data set. HC: home care; LR: logistic regression; MLP: multilayer perceptron; XGBoost: extreme gradient boosting.

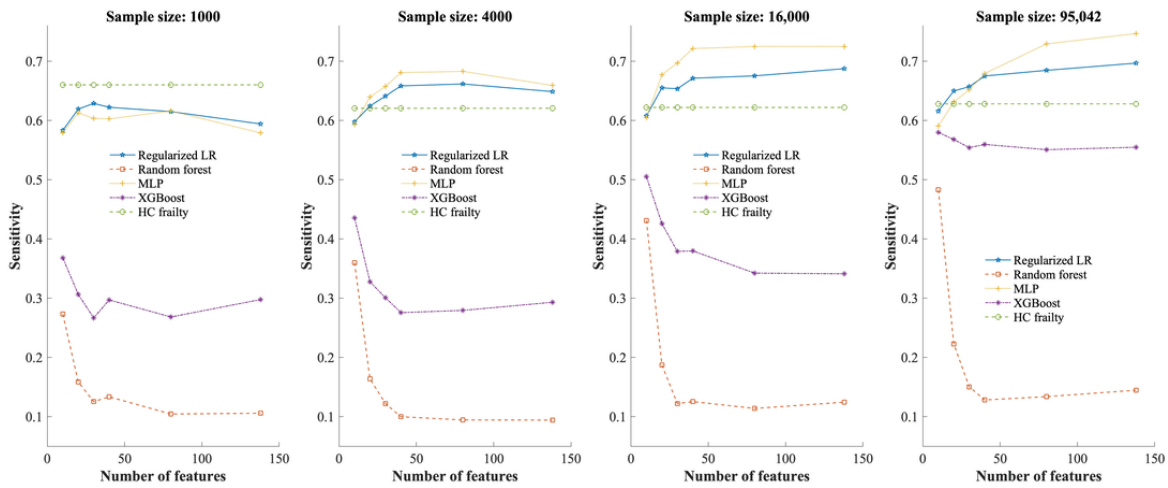


Figure 4. Average specificities of classifiers and frailty scale for 6-month mortality prediction on balanced data set. HC: home care; LR: logistic regression; MLP: multilayer perceptron; XGBoost: extreme gradient boosting.

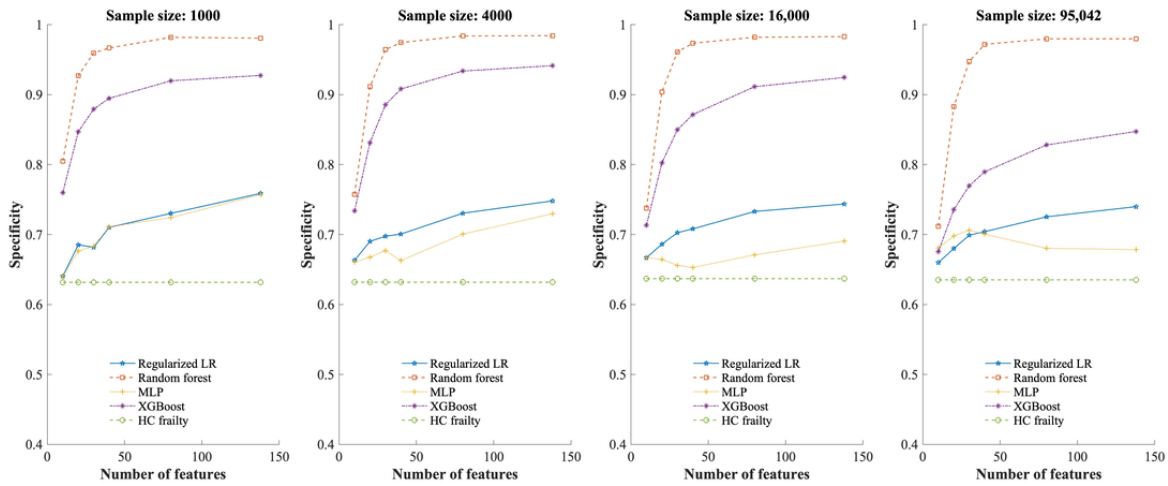
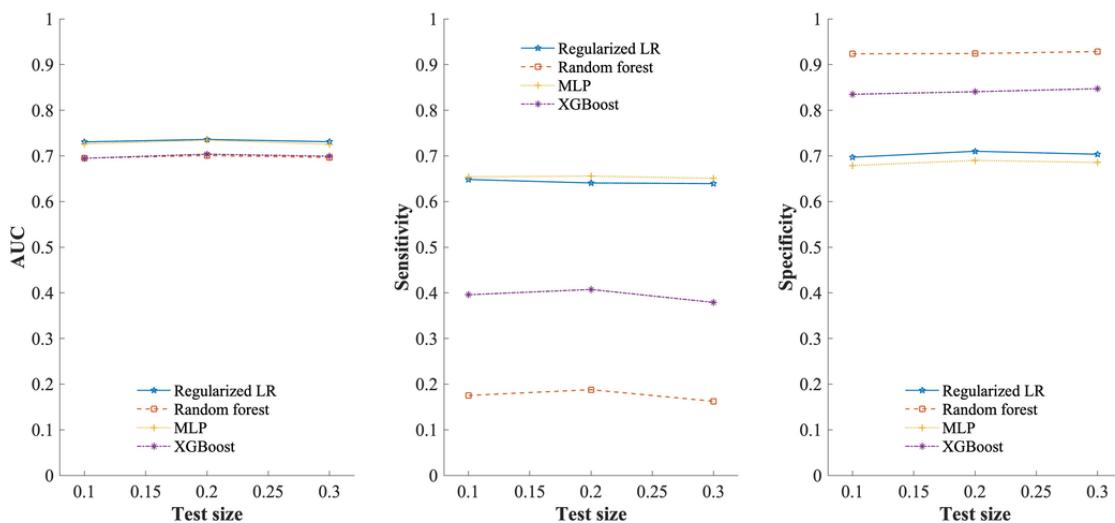


Figure 5. Average AUCs, sensitivities, and specificities of frailty scales for 6-month mortality prediction by test sizes on balanced data set. AUC: area under the curve; LR: logistic regression; MLP: multilayer perceptron; XGBoost: extreme gradient boosting.



Discussion

Principal Findings

In this retrospective study of older adults with the mandated standardized interRAI-HC assessment in New Zealand, we performed a series of simulations to evaluate the role of machine learning classifiers, features, and sample sizes on mortality prediction in identifying frail older individuals. We found that in most scenarios, particularly when dealing with large sample sizes and large numbers of features, 4 classifiers demonstrated significantly higher AUCs and sensitivities compared to the interRAI-HC Frailty Scale. All classifiers showed higher average specificities than the interRAI-HC Frailty Scale across all scenarios. Our simulation results showed that the predictive performance differed significantly by using different numbers of randomly selected features, varied sample sizes, and performance measures. Compared to machine learning classifiers, that is, RF, MLP, and XGBoost, logistic regressions provided higher average AUCs on 6-month mortality prediction when the number of features and sample sizes were not excessive. Even with a high number of features and very large samples, only slight improvements in average AUCs were observed in MLP and XGBoost. However, when the number of features and sample sizes were large, MLP demonstrated superior sensitivity, whereas RF exhibited superior specificity.

Interpretation in the Light of the Published Literature

In recent years, machine learning techniques have started to be used in various large-scale health care data sets to develop predictive algorithms for various adverse health outcomes, including hospitalization, mortality, and frailty in different populations [18,20,24,56]. For example, a recent study showed that by using only 10 or 11 features and 592 study subjects, the machine learning classifier support vector machines identified frail older adults with over 75% accuracy [45]. Another study also showed that by using 16 features, the machine learning classifier gradient boosting achieved 90% AUC on 30-day mortality prediction in patients with heart failure [19]. However, due to limitations in sample size and the number of available features, no study has systematically examined the role of methodological and database factors in the performance of various machine learning techniques. To our knowledge, our study is the first to use high-quality health care data of older adults receiving home care to investigate the performance of machine learning classifiers in identifying frail persons compared to an existing clinical scale and conventional logistic regressions. It is also the first to elucidate to what extent the performance is associated with the choice of classifier, sample size, and the number of features.

Contrary to our hypothesis, the application of machine learning classifiers did not improve the performance of mortality prediction for identifying frail older adults, as evaluated by AUC. This finding indicates that regularized logistic regression can perform sufficiently well and save computational resources when a well-structured, high-quality data source is used. One possible explanation for this result could be the nature of the features, as most of the items used to identify frail older adults are binary. Another reason may be the high reliability of

interRAI-HC data [21,57]. In a previous study that also used machine learning to predict frailty status, logistic regression demonstrated comparable or higher performance in various scenarios [27]. This previous study suggested that the tree-based classifiers performed better if the data set was of low quality and contained bad features, and that MLP could generally show a greater performance if the data set is large enough and has complex structure with many layers. In our study, the reason why MLP did not show superior performance on average AUCs could be due to only 1 hidden layer being used.

On the other hand, when the number of features and sample sizes were large, machine learning models demonstrated better performance than logistic regression on both sensitivity and specificity. Specifically, MLP exhibited superior sensitivity, which means that it was more effective at accurately identifying frail older adults receiving home care and were at high risk of adverse health outcomes. In contrast, RF demonstrated superior specificity, which means that it was better at correctly identifying those who were not at high risk of adverse health outcomes. In the context of frailty, where interventions such as exercise, reduction of polypharmacy, and adequate nutrition can attenuate and even reverse the condition [52], high sensitivity is considered more important than high specificity if a trade-off between the 2 measures is required.

Our study revealed that the RF and XGBoost classifiers had significantly lower sensitivities and higher specificities than logistic regression, while MLP had higher sensitivities and lower specificities. This finding is consistent with previous studies on identifying frailty. For example, a study using various machine learning methods to develop predictive models for frailty conditions in older individuals based on an administrative health database [18] observed lower sensitivities and higher specificities for RF when predicting urgent hospitalization, and higher sensitivities and lower specificities for MLP when predicting various health outcomes, including mortality, fracture, and preventable hospitalization. Another similar study that developed a validated case definition of frailty using machine learning classifiers [27] found significantly lower sensitivities and higher specificities for XGBoost and RF compared to logistic regression on balanced data using the default threshold. These findings collectively suggest that identifying frailty using machine learning techniques remains challenging and future research is warranted to investigate the performance of machine learning models in other populations and care settings.

Implications for Research, Policy, and Practice

We did not identify any machine learning classifier that performed consistently better than the others. The best classifier differed across experimental conditions. Our results demonstrate that the advantages of using machine learning techniques to identify frail older adults become more apparent as the sample size and number of features increase. The logistic regression demonstrated higher or comparable AUC compared to machine learning classifiers in most scenarios. This differs from previous studies that show that machine learning classifiers outperformed logistic regression or its variants in predicting adverse health outcomes [18,20,24-26]. With a sample size of 95,042 and 138 features, Ridge logistic regression achieved an average AUC

of 0.77 for 12-month mortality prediction. A logistic regression-based model developed by a previous study using interRAI-HC assessments of older persons in the New Zealand cohort targeting older individuals with complex comorbidities achieved an average AUC slightly higher (<0.01) than our result for 12-month mortality prediction [36]. The previous study used a slightly larger sample size of 104,436 and used a feature selection process to include only the features contributing over 1% to the performance. This may imply that a larger sample size and a feature selection process could further improve the predictive performance of logistic regression.

Strengths and Limitations

Our study used data collected from the interRAI instruments, standardized assessment instruments that have been developed by a collaborative network of health care professionals [21]. The interRAI instruments have been adopted in several jurisdictions to improve the quality of care for long-term care recipients, including Canada, Finland, Belgium, Italy, and Hong Kong. Therefore, the findings from this study may inform the identification of frail older adults for early interventions in similar care settings using interRAI assessments.

Our study has limitations. First, a successful measure of frailty should demonstrate satisfactory criterion validity against various adverse outcomes such as mortality, disability, hospitalization, and nursing home placement. Our study considered only mortality; therefore, it did not examine the accuracy of machine learning algorithms in predicting other adverse outcomes. Furthermore, we considered only 6- and 12-month mortality, resulting in an imbalanced data set that may yield higher specificity when using machine learning algorithms. It is also unclear whether the results can be extrapolated to other time intervals, such as 2 and 3 years. Further studies are needed to evaluate the prediction power of frailty against other critical outcomes. Second, the samples used in this study were limited to older adults receiving home care in New Zealand and most

participants were Europeans. Future studies are warranted to assess the generalizability of this study's findings. Third, we applied only 3 machine learning classifiers, chosen because they demonstrated better performance in several previous studies. The performance of other machine learning algorithms compared to regularized logistic regression was not investigated. Therefore, our conclusions are limited to the 3 algorithms examined. Fourth, calibration was not performed when training a machine learning classifier due to its additional computational costs, which may have affected the evaluation of model performance. The purpose of this study is to examine the impact of sample size and feature selection on the overall performance of each classifier in identifying frailty in older adults, rather than focusing on probability estimation or the quality of explanations provided by each model. It is worth noting that a recently published study [58] found that uncalibrated RF and XGBoost models performed similarly or even better than calibrated models in terms of accuracy and AUC. Therefore, the impact of calibration on our findings may not be severe. Finally, comparing the main features that affect the performance of different algorithms may improve the understanding of the construct of frailty. However, since the features in our simulation design were randomly selected across 1000 replications, the most important features identified from each run-in condition were not directly comparable. Therefore, we did not carry out further investigation on feature importance under different conditions.

Conclusions

Machine learning classifiers demonstrate considerable variability in prediction performance when assessed using different metrics. Regularized logistic regression is a reliable model for identifying frail older adults receiving home care, as indicated by the AUC, especially when the number of features and sample sizes are not excessively large. Conversely, MLP shows superior sensitivity, while RF demonstrates superior specificity when the number of features and sample sizes is large.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Supplementary experiments: features and results.

[DOCX File, 2598 KB - ai_v3i1e44185_app1.docx]

References

1. Kulmala J, Nykänen I, Hartikainen S. Frailty as a predictor of all-cause mortality in older men and women. *Geriatr Gerontol Int* 2014;14(4):899-905. [doi: [10.1111/ggi.12190](https://doi.org/10.1111/ggi.12190)] [Medline: [24666801](https://pubmed.ncbi.nlm.nih.gov/24666801/)]
2. Rodríguez-Mañas L, Féart C, Mann G, Viña J, Chatterji S, Chodsko-Zajko W, et al. Searching for an operational definition of frailty: a Delphi method based consensus statement: the frailty operative definition-consensus conference project. *J Gerontol A Biol Sci Med Sci* 2013;68(1):62-67 [FREE Full text] [doi: [10.1093/gerona/gls119](https://doi.org/10.1093/gerona/gls119)] [Medline: [22511289](https://pubmed.ncbi.nlm.nih.gov/22511289/)]
3. Romero-Ortuno R, Walsh CD, Lawlor BA, Kenny RA. A frailty instrument for primary care: findings from the Survey of Health, Ageing and Retirement in Europe (SHARE). *BMC Geriatr* 2010;10:57 [FREE Full text] [doi: [10.1186/1471-2318-10-57](https://doi.org/10.1186/1471-2318-10-57)] [Medline: [20731877](https://pubmed.ncbi.nlm.nih.gov/20731877/)]
4. Hubbard RE, Peel NM, Samanta M, Gray LC, Fries BE, Mitnitski A, et al. Derivation of a frailty index from the interRAI acute care instrument. *BMC Geriatr* 2015;15:27 [FREE Full text] [doi: [10.1186/s12877-015-0026-z](https://doi.org/10.1186/s12877-015-0026-z)] [Medline: [25887105](https://pubmed.ncbi.nlm.nih.gov/25887105/)]

5. Kaehr E, Visvanathan R, Malmstrom TK, Morley JE. Frailty in nursing homes: the FRAIL-NH Scale. *J Am Med Dir Assoc* 2015;16(2):87-89. [doi: [10.1016/j.jamda.2014.12.002](https://doi.org/10.1016/j.jamda.2014.12.002)] [Medline: [25556303](https://pubmed.ncbi.nlm.nih.gov/25556303/)]
6. Sobhani A, Fadayevatan R, Sharifi F, Kamrani AA, Ejtahed H, Hosseini RS, et al. The conceptual and practical definitions of frailty in older adults: a systematic review. *J Diabetes Metab Disord* 2021;20(2):1975-2013 [FREE Full text] [doi: [10.1007/s40200-021-00897-x](https://doi.org/10.1007/s40200-021-00897-x)] [Medline: [34900836](https://pubmed.ncbi.nlm.nih.gov/34900836/)]
7. Fried LP, Tangen CM, Walston J, Newman AB, Hirsch C, Gottdiener J, et al. Frailty in older adults: evidence for a phenotype. *J Gerontol A Biol Sci Med Sci* 2001;56(3):M146-M156 [FREE Full text] [doi: [10.1093/gerona/56.3.m146](https://doi.org/10.1093/gerona/56.3.m146)] [Medline: [11253156](https://pubmed.ncbi.nlm.nih.gov/11253156/)]
8. Xue QL. The frailty syndrome: definition and natural history. *Clin Geriatr Med* 2011;27(1):1-15 [FREE Full text] [doi: [10.1016/j.cger.2010.08.009](https://doi.org/10.1016/j.cger.2010.08.009)] [Medline: [21093718](https://pubmed.ncbi.nlm.nih.gov/21093718/)]
9. Kojima G, Iliffe S, Walters K. Frailty index as a predictor of mortality: a systematic review and meta-analysis. *Age Ageing* 2018;47(2):193-200 [FREE Full text] [doi: [10.1093/ageing/afx162](https://doi.org/10.1093/ageing/afx162)] [Medline: [29040347](https://pubmed.ncbi.nlm.nih.gov/29040347/)]
10. Fried LP, Cohen AA, Xue QL, Walston J, Bandeen-Roche K, Varadhan R. The physical frailty syndrome as a transition from homeostatic symphony to cacophony. *Nat Aging* 2021;1(1):36-46 [FREE Full text] [doi: [10.1038/s43587-020-00017-z](https://doi.org/10.1038/s43587-020-00017-z)] [Medline: [34476409](https://pubmed.ncbi.nlm.nih.gov/34476409/)]
11. Rockwood K, Mitnitski A. Frailty in relation to the accumulation of deficits. *J Gerontol A Biol Sci Med Sci* 2007;62(7):722-727 [FREE Full text] [doi: [10.1093/gerona/62.7.722](https://doi.org/10.1093/gerona/62.7.722)] [Medline: [17634318](https://pubmed.ncbi.nlm.nih.gov/17634318/)]
12. Rockwood K. What would make a definition of frailty successful? *Age Ageing* 2005;34(5):432-434 [FREE Full text] [doi: [10.1093/ageing/afi146](https://doi.org/10.1093/ageing/afi146)] [Medline: [16107450](https://pubmed.ncbi.nlm.nih.gov/16107450/)]
13. Mitnitski AB, Mogilner AJ, Rockwood K. Accumulation of deficits as a proxy measure of aging. *ScientificWorldJournal* 2001;1:323-336 [FREE Full text] [doi: [10.1100/tsw.2001.58](https://doi.org/10.1100/tsw.2001.58)] [Medline: [12806071](https://pubmed.ncbi.nlm.nih.gov/12806071/)]
14. Makizako H, Shimada H, Tsutsumimoto K, Lee S, Doi T, Nakakubo S, et al. Social frailty in community-dwelling older adults as a risk factor for disability. *J Am Med Dir Assoc* 2015;16(11):1003.e7-1003.e11. [doi: [10.1016/j.jamda.2015.08.023](https://doi.org/10.1016/j.jamda.2015.08.023)] [Medline: [26482055](https://pubmed.ncbi.nlm.nih.gov/26482055/)]
15. Teo N, Gao Q, Nyunt MSZ, Wee SL, Ng TP. Social frailty and functional disability: findings from the Singapore longitudinal ageing studies. *J Am Med Dir Assoc* 2017;18(7):637.e13-637.e19. [doi: [10.1016/j.jamda.2017.04.015](https://doi.org/10.1016/j.jamda.2017.04.015)] [Medline: [28648903](https://pubmed.ncbi.nlm.nih.gov/28648903/)]
16. Bunt S, Steverink N, Olthof J, van der Schans CP, Hobbelen JSM. Social frailty in older adults: a scoping review. *Eur J Ageing* 2017;14(3):323-334 [FREE Full text] [doi: [10.1007/s10433-017-0414-7](https://doi.org/10.1007/s10433-017-0414-7)] [Medline: [28936141](https://pubmed.ncbi.nlm.nih.gov/28936141/)]
17. Ravaglia G, Forti P, Lucicesare A, Pisacane N, Rietti E, Patterson C. Development of an easy prognostic score for frailty outcomes in the aged. *Age Ageing* 2008;37(2):161-166 [FREE Full text] [doi: [10.1093/ageing/afm195](https://doi.org/10.1093/ageing/afm195)] [Medline: [18238805](https://pubmed.ncbi.nlm.nih.gov/18238805/)]
18. Tarekegn A, Ricceri F, Costa G, Ferracin E, Giacobini M. Predictive modeling for frailty conditions in elderly people: machine learning approaches. *JMIR Med Inform* 2020;8(6):e16678 [FREE Full text] [doi: [10.2196/16678](https://doi.org/10.2196/16678)] [Medline: [32442149](https://pubmed.ncbi.nlm.nih.gov/32442149/)]
19. Ju C, Zhou J, Lee S, Tan MS, Liu T, Bazoukis G, et al. Derivation of an electronic frailty index for predicting short-term mortality in heart failure: a machine learning approach. *ESC Heart Fail* 2021;8(4):2837-2845 [FREE Full text] [doi: [10.1002/ehf2.13358](https://doi.org/10.1002/ehf2.13358)] [Medline: [34080784](https://pubmed.ncbi.nlm.nih.gov/34080784/)]
20. Parikh RB, Manz C, Chivers C, Regli SH, Braun J, Draugelis ME, et al. Machine learning approaches to predict 6-month mortality among patients with cancer. *JAMA Netw Open* 2019;2(10):e1915997 [FREE Full text] [doi: [10.1001/jamanetworkopen.2019.15997](https://doi.org/10.1001/jamanetworkopen.2019.15997)] [Medline: [31651973](https://pubmed.ncbi.nlm.nih.gov/31651973/)]
21. Hirdes JP, Ljunggren G, Morris JN, Frijters DHM, Soveri HF, Gray L, et al. Reliability of the interRAI suite of assessment instruments: a 12-country study of an integrated health information system. *BMC Health Serv Res* 2008;8:277 [FREE Full text] [doi: [10.1186/1472-6963-8-277](https://doi.org/10.1186/1472-6963-8-277)] [Medline: [19115991](https://pubmed.ncbi.nlm.nih.gov/19115991/)]
22. Murdoch TB, Detsky AS. The inevitable application of big data to health care. *JAMA* 2013;309(13):1351-1352. [doi: [10.1001/jama.2013.393](https://doi.org/10.1001/jama.2013.393)] [Medline: [23549579](https://pubmed.ncbi.nlm.nih.gov/23549579/)]
23. Fragidis LL, Chatzoglou PD. Implementation of a nationwide electronic health record (EHR). *Int J Health Care Qual Assur* 2018;31(2):116-130. [doi: [10.1108/IJHCQA-09-2016-0136](https://doi.org/10.1108/IJHCQA-09-2016-0136)] [Medline: [29504871](https://pubmed.ncbi.nlm.nih.gov/29504871/)]
24. Vaid A, Somani S, Russak AJ, De Freitas JK, Chaudhry FF, Paranjpe I, et al. Machine learning to predict mortality and critical events in a cohort of patients with COVID-19 in New York City: model development and validation. *J Med Internet Res* 2020;22(11):e24018 [FREE Full text] [doi: [10.2196/24018](https://doi.org/10.2196/24018)] [Medline: [33027032](https://pubmed.ncbi.nlm.nih.gov/33027032/)]
25. Liu J, Wu J, Liu S, Li M, Hu K, Li K. Predicting mortality of patients with acute kidney injury in the ICU using XGBoost model. *PLoS One* 2021;16(2):e0246306 [FREE Full text] [doi: [10.1371/journal.pone.0246306](https://doi.org/10.1371/journal.pone.0246306)] [Medline: [33539390](https://pubmed.ncbi.nlm.nih.gov/33539390/)]
26. Lin K, Hu Y, Kong G. Predicting in-hospital mortality of patients with acute kidney injury in the ICU using random forest model. *Int J Med Inform* 2019;125:55-61. [doi: [10.1016/j.ijmedinf.2019.02.002](https://doi.org/10.1016/j.ijmedinf.2019.02.002)] [Medline: [30914181](https://pubmed.ncbi.nlm.nih.gov/30914181/)]
27. Aponte-Hao S, Wong ST, Thandi M, Ronksley P, McBrien K, Lee J, et al. Machine learning for identification of frailty in Canadian primary care practices. *Int J Popul Data Sci* 2021;6(1):1650 [FREE Full text] [doi: [10.23889/ijpds.v6i1.1650](https://doi.org/10.23889/ijpds.v6i1.1650)] [Medline: [34541337](https://pubmed.ncbi.nlm.nih.gov/34541337/)]
28. Morris JN, Howard EP, Steel KR. Development of the interRAI home care frailty scale. *BMC Geriatr* 2016;16(1):188 [FREE Full text] [doi: [10.1186/s12877-016-0364-5](https://doi.org/10.1186/s12877-016-0364-5)] [Medline: [27871235](https://pubmed.ncbi.nlm.nih.gov/27871235/)]

29. Hirdes JP, van Everdingen C, Ferris J, Franco-Martin M, Fries BE, Heikkilä J, et al. The interRAI suite of mental health assessment instruments: an integrated system for the continuum of care. *Front Psychiatry* 2020;10:926 [FREE Full text] [doi: [10.3389/fpsy.2019.00926](https://doi.org/10.3389/fpsy.2019.00926)] [Medline: [32076412](https://pubmed.ncbi.nlm.nih.gov/32076412/)]
30. Corsonello A, Lattanzio F, Pedone C, Garasto S, Laino I, Bustacchini S, et al. Prognostic significance of the short physical performance battery in older patients discharged from acute care hospitals. *Rejuvenation Res* 2012;15(1):41-48 [FREE Full text] [doi: [10.1089/rej.2011.1215](https://doi.org/10.1089/rej.2011.1215)] [Medline: [22004280](https://pubmed.ncbi.nlm.nih.gov/22004280/)]
31. Afilalo J, Lauck S, Kim DH, Lefèvre T, Piazza N, Lachapelle K, et al. Frailty in older adults undergoing aortic valve replacement: the FRAILTY-AVR study. *J Am Coll Cardiol* 2017;70(6):689-700 [FREE Full text] [doi: [10.1016/j.jacc.2017.06.024](https://doi.org/10.1016/j.jacc.2017.06.024)] [Medline: [28693934](https://pubmed.ncbi.nlm.nih.gov/28693934/)]
32. Campo G, Maietti E, Tonet E, Biscaglia S, Ariza-Solè A, Pavasini R, et al. The assessment of scales of frailty and physical performance improves prediction of major adverse cardiac events in older adults with acute coronary syndrome. *J Gerontol A Biol Sci Med Sci* 2020;75(6):1113-1119 [FREE Full text] [doi: [10.1093/gerona/glz123](https://doi.org/10.1093/gerona/glz123)] [Medline: [31075167](https://pubmed.ncbi.nlm.nih.gov/31075167/)]
33. Espauella J, Arnau A, Cubí D, Amblàs J, Yáñez A. Time-dependent prognostic factors of 6-month mortality in frail elderly patients admitted to post-acute care. *Age Ageing* 2007;36(4):407-413 [FREE Full text] [doi: [10.1093/ageing/afm033](https://doi.org/10.1093/ageing/afm033)] [Medline: [17395620](https://pubmed.ncbi.nlm.nih.gov/17395620/)]
34. Abey-Nesbit R, Bergler U, Pickering JW, Nishtala PS, Jamieson H. Development and validation of a frailty index compatible with three interRAI assessment instruments. *Age Ageing* 2022;51(8):afac178 [FREE Full text] [doi: [10.1093/ageing/afac178](https://doi.org/10.1093/ageing/afac178)] [Medline: [35930721](https://pubmed.ncbi.nlm.nih.gov/35930721/)]
35. Kerminen H, Huhtala H, Jäntti P, Valvanne J, Jämsen E. Frailty index and functional level upon admission predict hospital outcomes: an interRAI-based cohort study of older patients in post-acute care hospitals. *BMC Geriatr* 2020;20(1):160 [FREE Full text] [doi: [10.1186/s12877-020-01550-7](https://doi.org/10.1186/s12877-020-01550-7)] [Medline: [32370740](https://pubmed.ncbi.nlm.nih.gov/32370740/)]
36. Pickering JW, Abey-Nesbit R, Allore H, Jamieson H. Development and validation of multivariable mortality risk-prediction models in older people undergoing an interRAI home-care assessment (RiskOP). *EClinicalMedicine* 2020;29-30:100614 [FREE Full text] [doi: [10.1016/j.eclinm.2020.100614](https://doi.org/10.1016/j.eclinm.2020.100614)] [Medline: [33437945](https://pubmed.ncbi.nlm.nih.gov/33437945/)]
37. Sternberg SA, Schwartz AW, Karunanathan S, Bergman H, Clarfield AM. The identification of frailty: a systematic literature review. *J Am Geriatr Soc* 2011;59(11):2129-2138. [doi: [10.1111/j.1532-5415.2011.03597.x](https://doi.org/10.1111/j.1532-5415.2011.03597.x)] [Medline: [22091630](https://pubmed.ncbi.nlm.nih.gov/22091630/)]
38. Biau G. Analysis of a random forests model. *J Mach Learn Res* 2012;13:1063-1095 [FREE Full text]
39. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. New York, NY, US: Association for Computing Machinery; 2016 Presented at: Kdd'16: Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining; August 13-17, 2016; San Francisco, California, USA p. 785-794. [doi: [10.1145/2939672.2939785](https://doi.org/10.1145/2939672.2939785)]
40. Colombet I, Ruelland A, Chatellier G, Gueyffier F, Degoulet P, Jaurent M. Models to predict cardiovascular risk: comparison of CART, multilayer perceptron and logistic regression. *Proc AMIA Symp* 2000:156-160 [FREE Full text] [Medline: [11079864](https://pubmed.ncbi.nlm.nih.gov/11079864/)]
41. Tibshirani R. Regression shrinkage and selection via the lasso. *J R Stat Soc, Ser B, Methodol* 1996;58(1):267-288. [doi: [10.1111/j.2517-6161.1996.tb02080.x](https://doi.org/10.1111/j.2517-6161.1996.tb02080.x)]
42. Hoerl AE, Kennard RW. Ridge regression: biased estimation for nonorthogonal problems. *Technometrics* 1970;12(1):55-67. [doi: [10.1080/00401706.1970.10488634](https://doi.org/10.1080/00401706.1970.10488634)]
43. Altukhova O. Choice of method imputation missing values for obstetrics clinical data. *Procedia Comput Sci* 2020;176:976-984 [FREE Full text] [doi: [10.1016/j.procs.2020.09.093](https://doi.org/10.1016/j.procs.2020.09.093)]
44. Viloría A, Pineda Lezama OB, Mercado-Caruzo N. Unbalanced data processing using oversampling: machine learning. *Procedia Comput Sci* 2020;175:108-113 [FREE Full text] [doi: [10.1016/j.procs.2020.07.018](https://doi.org/10.1016/j.procs.2020.07.018)]
45. Ambagtsheer RC, Shafiabady N, Dent E, Seiboth C, Beilby J. The application of artificial intelligence (AI) techniques to identify frailty within a residential aged care administrative data set. *Int J Med Inform* 2020;136:104094. [doi: [10.1016/j.ijmedinf.2020.104094](https://doi.org/10.1016/j.ijmedinf.2020.104094)] [Medline: [32058264](https://pubmed.ncbi.nlm.nih.gov/32058264/)]
46. Williamson T, Aponte-Hao S, Mele B, Lethebe BC, Leduc C, Thandi M, et al. Developing and validating a primary care EMR-based frailty definition using machine learning. *Int J Popul Data Sci* 2020;5(1):1344 [FREE Full text] [doi: [10.23889/ijpds.v5i1.1344](https://doi.org/10.23889/ijpds.v5i1.1344)] [Medline: [32935059](https://pubmed.ncbi.nlm.nih.gov/32935059/)]
47. Powers DMW. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. *J Mach Learn Technol* 2011;2:37-63 [FREE Full text]
48. Kiely DK, Cupples LA, Lipsitz LA. Validation and comparison of two frailty indexes: the MOBILIZE Boston study. *J Am Geriatr Soc* 2009;57(9):1532-1539 [FREE Full text] [doi: [10.1111/j.1532-5415.2009.02394.x](https://doi.org/10.1111/j.1532-5415.2009.02394.x)] [Medline: [19682112](https://pubmed.ncbi.nlm.nih.gov/19682112/)]
49. Hadanny A, Shouval R, Wu J, Gale CP, Unger R, Zahger D, et al. Machine learning-based prediction of 1-year mortality for acute coronary syndrome. *J Cardiol* 2022;79(3):342-351 [FREE Full text] [doi: [10.1016/j.jicc.2021.11.006](https://doi.org/10.1016/j.jicc.2021.11.006)] [Medline: [34857429](https://pubmed.ncbi.nlm.nih.gov/34857429/)]
50. Taylor RA, Pare JR, Venkatesh AK, Mowafi H, Melnick ER, Fleischman W, et al. Prediction of in-hospital mortality in emergency department patients with sepsis: a local big data-driven, machine learning approach. *Acad Emerg Med* 2016;23(3):269-278 [FREE Full text] [doi: [10.1111/acem.12876](https://doi.org/10.1111/acem.12876)] [Medline: [26679719](https://pubmed.ncbi.nlm.nih.gov/26679719/)]
51. Fawcett T. An introduction to ROC analysis. *Pattern Recognit Lett* 2006;27(8):861-874. [doi: [10.1016/j.patrec.2005.10.010](https://doi.org/10.1016/j.patrec.2005.10.010)]

52. Morley JE, Vellas B, van Kan GA, Anker SD, Bauer JM, Bernabei R, et al. Frailty consensus: a call to action. *J Am Med Dir Assoc* 2013;14(6):392-397 [FREE Full text] [doi: [10.1016/j.jamda.2013.03.022](https://doi.org/10.1016/j.jamda.2013.03.022)] [Medline: [23764209](https://pubmed.ncbi.nlm.nih.gov/23764209/)]
53. Sasaki Y. The truth of the F-measure. *Teach tutor mater* 2007;1(5):1-5 [FREE Full text]
54. Accuracy (trueness and precision) of measurement methods and results. ISO. 1998. URL: <https://www.iso.org/standard/79066.html> [accessed 2024-01-09]
55. Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS One* 2015;10(3):e0118432 [FREE Full text] [doi: [10.1371/journal.pone.0118432](https://doi.org/10.1371/journal.pone.0118432)] [Medline: [25738806](https://pubmed.ncbi.nlm.nih.gov/25738806/)]
56. Jones A, Costa AP, Pesevski A, McNicholas PD. Predicting hospital and emergency department utilization among community-dwelling older adults: statistical and machine learning approaches. *PLoS One* 2018;13(11):e0206662 [FREE Full text] [doi: [10.1371/journal.pone.0206662](https://doi.org/10.1371/journal.pone.0206662)] [Medline: [30383850](https://pubmed.ncbi.nlm.nih.gov/30383850/)]
57. Hogeveen SE, Chen J, Hirdes JP. Evaluation of data quality of interRAI assessments in home and community care. *BMC Med Inform Decis Mak* 2017;17(1):150 [FREE Full text] [doi: [10.1186/s12911-017-0547-9](https://doi.org/10.1186/s12911-017-0547-9)] [Medline: [29084534](https://pubmed.ncbi.nlm.nih.gov/29084534/)]
58. Löfström H, Löfström T, Johansson U, Sönströd C. Investigating the impact of calibration on the quality of explanations. *Ann Math Artif Intell* 2023:1-18 [FREE Full text] [doi: [10.1007/s10472-023-09837-2](https://doi.org/10.1007/s10472-023-09837-2)]

Abbreviations

AUC: area under the curve
CHS: Cardiovascular Health Study
HC: home care
interRAI-HC: interRAI-Home Care
LASSO: Least Absolute Shrinkage and Selection Operator
MLP: multilayer perceptron
RF: random forest
XGBoost: extreme gradient boosting

Edited by K El Emam, B Malin; submitted 09.11.22; peer-reviewed by C Bian, JR Medina, D Han; comments to author 02.07.23; revised version received 22.07.23; accepted 01.01.24; published 31.01.24.

Please cite as:

Pan C, Luo H, Cheung G, Zhou H, Cheng R, Cullum S, Wu C

Identifying Frailty in Older Adults Receiving Home Care Assessment Using Machine Learning: Longitudinal Observational Study on the Role of Classifier, Feature Selection, and Sample Size

JMIR AI 2024;3:e44185

URL: <https://ai.jmir.org/2024/1/e44185>

doi: [10.2196/44185](https://doi.org/10.2196/44185)

PMID:

©Cheng Pan, Hao Luo, Gary Cheung, Huiquan Zhou, Reynold Cheng, Sarah Cullum, Chuan Wu. Originally published in JMIR AI (<https://ai.jmir.org>), 31.01.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR AI, is properly cited. The complete bibliographic information, a link to the original publication on <https://www.ai.jmir.org/>, as well as this copyright and license information must be included.

Original Paper

Development of Lung Cancer Risk Prediction Machine Learning Models for Equitable Learning Health System: Retrospective Study

Anjun Chen¹, PhD; Erman Wu², MD; Ran Huang², MS; Bairong Shen², PhD; Ruobing Han³, MA; Jian Wen⁴, PhD; Zhiyong Zhang¹, PhD; Qinghua Li⁴, MD, PhD

¹School of Public Health, Guilin Medical University, Guilin, China

²West China Hospital, Chengdu, China

³Guilin Medical University, Guilin, China

⁴Department of Neurology, Guilin Medical University Affiliated Hospital, Guilin, Guangxi, China

Corresponding Author:

Qinghua Li, MD, PhD

Department of Neurology

Guilin Medical University Affiliated Hospital

15 Lequn Road

Guilin, Guangxi, 541000

China

Phone: 86 15878361508

Email: qhli1999@glmc.edu.cn

Abstract

Background: A significant proportion of young at-risk patients and nonsmokers are excluded by the current guidelines for lung cancer (LC) screening, resulting in low-screening adoption. The vision of the US National Academy of Medicine to transform health systems into learning health systems (LHS) holds promise for bringing necessary structural changes to health care, thereby addressing the exclusivity and adoption issues of LC screening.

Objective: This study aims to realize the LHS vision by designing an equitable, machine learning (ML)-enabled LHS unit for LC screening. It focuses on developing an inclusive and practical LC risk prediction model, suitable for initializing the ML-enabled LHS (ML-LHS) unit. This model aims to empower primary physicians in a clinical research network, linking central hospitals and rural clinics, to routinely deliver risk-based screening for enhancing LC early detection in broader populations.

Methods: We created a standardized data set of health factors from 1397 patients with LC and 1448 control patients, all aged 30 years and older, including both smokers and nonsmokers, from a hospital's electronic medical record system. Initially, a data-centric ML approach was used to create inclusive ML models for risk prediction from all available health factors. Subsequently, a quantitative distribution of LC health factors was used in feature engineering to refine the models into a more practical model with fewer variables.

Results: The initial inclusive 250-variable XGBoost model for LC risk prediction achieved performance metrics of 0.86 recall, 0.90 precision, and 0.89 accuracy. Post feature refinement, a practical 29-variable XGBoost model was developed, displaying performance metrics of 0.80 recall, 0.82 precision, and 0.82 accuracy. This model met the criteria for initializing the ML-LHS unit for risk-based, inclusive LC screening within clinical research networks.

Conclusions: This study designed an innovative ML-LHS unit for a clinical research network, aiming to sustainably provide inclusive LC screening to all at-risk populations. It developed an inclusive and practical XGBoost model from hospital electronic medical record data, capable of initializing such an ML-LHS unit for community and rural clinics. The anticipated deployment of this ML-LHS unit is expected to significantly improve LC-screening rates and early detection among broader populations, including those typically overlooked by existing screening guidelines.

(JMIR AI 2024;3:e56590) doi:[10.2196/56590](https://doi.org/10.2196/56590)

KEYWORDS

lung cancer; risk prediction; early detection; learning health system; LHS; machine learning; ML; artificial intelligence; AI; predictive model

Introduction

Lung Cancer–Screening Challenges

Lung cancer (LC) is the second most common cancer and the leading cause of cancer deaths worldwide [1]. It accounted for an estimated 2.2 million new cases and 1.8 million deaths in 2020. Screening for early detection of LC is a crucial strategy to combat this deadly disease [2]. LC-screening guidelines recommend that heavy smokers aged 50-80 years undergo LC screening [3]. Clinical trials have shown about a 20% reduction in LC mortality due to screening with low-dose computed tomography [4].

However, nonsmoking adults and individuals younger than 50 years are often excluded from LC-screening guidelines, despite representing a significant percentage of patients with LC worldwide [5,6]. Statistical risk prediction models, such as PLCOm2012, have been used to recommend LC screening for smokers [7]. The subsequent PLCOall2014 model included nonsmokers in risk evaluation [8], but its impact on screening uptake was unclear. In addition, the adoption of LC screening is low; for instance, only about 5% of the at-risk population in the United States has undergone LC screening [9].

There have been numerous research efforts to overcome these challenges, but their results were inconclusive and unsatisfactory [10]. Researchers have proposed individualized risk-based screening approaches for both smokers and nonsmokers [11]. In 2018, the PLCO model developer reviewed several traditional risk prediction models and suggested that the including biomarkers might help identify individuals who could benefit from LC screening [12]. The PanCan study demonstrated that selecting participants for LC screening based on risk modeling could identify patients with early-stage LC [13]. A recent systematic review concluded that further research is needed to optimize risk-based LC screening [14]. Concurrently, an updated evidence report for the US Preventive Services Task Force indicated that screening high-risk individuals with low-dose computed tomography could reduce LC mortality but might also lead to false positives, resulting in unnecessary tests and invasive procedures [15].

As electronic medical records (EMRs) become prevalent in hospitals, several machine learning (ML) models have been developed using EMR data for LC risk prediction. Kaiser researchers used a small set of preselected variables to identify patients with early-stage LC from routine clinical and laboratory data [16,17]. Stanford researchers developed an ML model to predict the 1-year risk of incident LC using more than 33,000 features from EMR data [18]. Deep learning with convolutional neural networks applied to EMR data from 2 million patients produced a high-performance LC risk prediction model [19]. However, the widespread deployment of these models for risk-based LC screening is yet to be determined.

The Learning Health System Approach

Over a decade ago, the US National Academy of Medicine (NAM) identified some major shortcomings in the current clinical evidence generation enterprise and proposed the vision of learning health systems (LHS) to address these issues [20-22].

First, many guidelines are primarily based on clinical trials with narrow scopes, failing to fully represent real-world scenarios. For instance, the exclusion of nonsmokers and younger populations from the LC guidelines might be a result of these narrow scopes. Second, the slow dissemination of evidence from discovery to clinical practice contributes to the low adoption rate of LC screening. To address these significant challenges, NAM envisions transforming health systems into LHS to bring necessary structural changes to health care. One of the most significant system-level changes in LHS is that embedding clinical research becomes into routine clinical delivery, facilitating more efficient generation of real-world evidence from real-world data (RWD) of patients and faster dissemination of new evidence to practices. Efficient evidence generation also necessitates innovations in clinical trial methodologies, such as pragmatic clinical trials [23,24].

We believe that NAM's LHS vision points in the right direction to address the exclusivity, bias, and adoption issues of LC screening. In pursuing sustainable, long-term solutions for inclusive screening and increased screening rates, we believe that system-level innovations are essential. We have focused on two interdependent considerations: (1) more inclusive intervention: exploring data-centric, risk-based LC-screening recommendations instead of blunt exclusions of certain demographic groups; and (2) broader access to the intervention: applying ML-based artificial intelligence (AI) to enable doctors in community and rural primary care to conduct routine risk-based LC screening. Our goal is to assess whether identifying at-risk individuals anywhere using the LHS approach can help close the gap in LC-screening disparities.

These considerations necessitate at least two innovations: (1) a new ML-enabled LHS unit that can continuously improve ML models and thus enhance risk prediction services. Our first ML-enabled LHS (ML-LHS) simulation study using synthetic patient data demonstrated performance improvement of LC risk prediction ML models over time [25]. (2) ML models that are inclusive in terms of patient populations and practical for use in low-resource clinics. Previously, by applying a data-centric EMR ML approach and feature engineering based on a quantitative distribution of health factors derived from EMR data [26], we successfully developed an inclusive and practical ML model for predicting the risk of nasopharyngeal cancer [27].

Aims

This study aimed to design an equitable ML-LHS unit for LC screening and to develop an inclusive and practical LC risk prediction model suitable for initializing the LC-screening ML-LHS unit. The future deployment of this new LC ML-LHS unit will aid in implementing risk-based LC screening across populations broader than those currently covered by existing LC-screening guidelines, thereby improving both patient coverage and LC-screening rate.

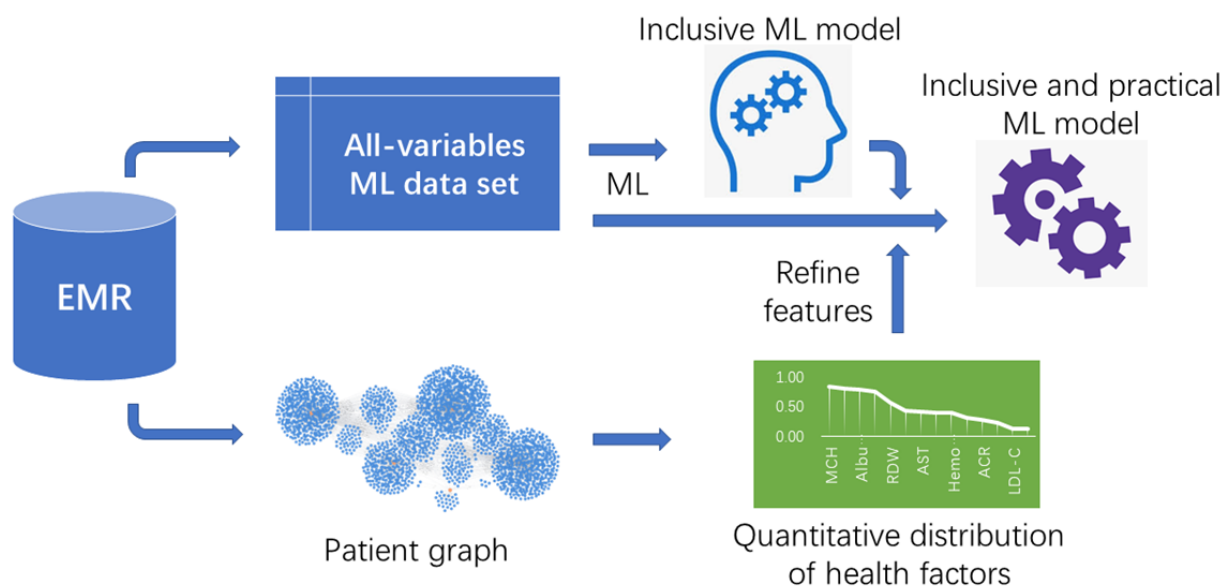
Methods

Hybrid EMR ML Pipeline for Inclusive and Practical LC ML Model

We designed a hybrid EMR ML pipeline to create an inclusive and practical ML model for LC risk prediction (see Figure 1). In step 1, data related to all health factors associated with LC are collected from the EMR. Common ML algorithms, such as

XGBoost, are then used to train risk prediction models using these data. In step 2, a patient graph is constructed using all health factors in the EMR, which produces a quantitative LC health factor distribution. In step 3, feature engineering, based on the health factor distribution, refines the model into a more practical one with fewer variables. The recently published patient graph analysis method is used to generate this quantitative distribution of health factors from hospital EMR data [26].

Figure 1. Hybrid EMR ML pipeline for developing inclusive and practical machine learning models for lung cancer risk prediction. The inclusive ML model uses as many health factor variables from EMR as possible. In contrast, the practical ML model uses a small number of variables that are readily available in low-resource clinics. The quantitative distribution of health factor distribution, derived from real-world patient data, aids in refining the features of the inclusive model to formulate the practical model. EMR: electronic medical record; ML: machine learning.



Standardized Patient Data Collection

Deidentified patient medical records were generated from the hospital's EMR and relevant databases, covering the period from January 2018 to June 2021. These data sets were securely stored on a data server managed by the hospital's informatics department. The data set encompassed about 1 million patients and 7 million outpatient and inpatient encounters. The records excluded all fields containing personal information, such as patient names, birth dates, personal IDs, contact details, and addresses. Original hospital identifiers for patients and encounters were replaced by random numbers, not linked to the patients.

Due to the absence of applicable codes for diagnoses in the EMR, Chinese synonyms for LC were used to identify patients with LC. The targeted data set included 1397 patients with LC aged 30 years and older. In addition, 1448 patients aged 30 years and older with no LC were randomly selected to form the background or control data set. We maintained similar numbers of patients in the target and control groups to preserve class balance. However, data standardization, being time-consuming, limited the number of patients in the final structured data set.

Based on our experience in building multiple models from EMR, the minimal number is approximately 1000 target patients and 1000 background patients.

Deidentified records of outpatient and inpatient visits, diagnoses, laboratory tests, and procedures were imported into a custom data collection tool on the data server. This tool automatically extracted laboratory test data for storage in a MongoDB database, provided by MongoDB Inc. Our researchers manually curated data from patient record texts and entered them into the database. Data were categorized into 9 categories: disease and condition, symptom, medical history, observation, laboratory test, procedure, medication, treatment, and other risk factors. To overcome the lack of coding and standardization in the records, practical rules were established to ensure consistency in data collection. Synonyms were automatically converted to local "standard terms" with corresponding local codes, culminating in local "standard data." For each patient with LC, only those data leading to the final diagnosis of LC were collected, forming a patient diagnosis journey (PDJ) object comprising 1 or multiple encounters. For each background patient, all encounters within the 3.5-years period were included. When exporting PDJ data to a comma-separated values file for

analysis, only the most recent data for each health factor in the PDJ were selected.

EMR ML for Inclusive LC Risk Prediction Models

All continuous numeric data in the profiles were converted to categorical data. For example, age ranges were established as 30-50, 50-70, and more than 70 years; drinking levels were categorized as 0-2, and >2 drinks per day; and smoking levels were divided into 0, 1-20, and >20 cigarettes per day. Laboratory test results had predefined categorized such as normal or abnormal, true or false, positive or negative, and high, medium, or low. After this conversion, profiles of patient with LC encompassed more than 58,000 data items and 2066 codes, while background patient profiles comprised more than 46,000 data items and 1298 codes. Subsequently, the profile data were structured into a horizontal table for ML, labeling patients with LC as “1” and background patients as “0.”

Codes were organized based on the number of associated patients with LC. Various sets of codes, exceeding a cutoff of 10 patients with LC, were selected by different criteria for ML. For the LC risk prediction study, all codes related to cancer diseases, procedures, medications, and treatments were omitted. In addition, diagnostic imaging procedures commonly used for patients with cancer but not for background patients were also excluded.

In developing ML models, we used the XGBoost Python library [28]. XGBoost is known for parallel tree boosting and its efficient management of missing data. The Python library scikit-learn from Scikit-learn.org was used for all other ML tasks [29]. The free Jupyter Notebook tool was used to conduct ML experiments [30]. The Pandas library was used for reading and writing comma-separated values files and manipulating data tables. The data set was divided into training (60%), tuning (20%), and validating (20%) subsets. Using the default hyperparameters, the XGBoost classifier was fitted with the training and tuning sets, and the resulting model was independently validated by the validation data set [31]. The model's effectiveness in risk prediction was evaluated using key metrics such as recall, precision, area under the receiver operating characteristic curve (AUROC), and accuracy. Receiver operating characteristic (ROC) curve and reliability (or calibration) curve were drawn by calling the corresponding Scikit-learn functions.

By comparing the performances of models built from different variable sets, an inclusive variable set was established. Using this set, XGBoost was compared with 3 other commonly used algorithms: random forest (RF), support vector machines (SVM), and k-nearest neighbors (KNN). These algorithms were executed using Scikit-learn classifiers with default parameters. The main reason for evaluating only the common algorithms is because they are promising in delivering the initial acceptable performance required by our LHS design, and their deployment is easier and cost-efficient. Only if this test fails will we test more complex algorithms like neural networks.

Building Practical ML Prediction Models

In the final refinement step of our hybrid ML pipeline, a quantitative distribution of LC health factors was generated

directly from the same EMR data through patient graph analysis [32]. In the patient graph, health factors are connected to patients with LC and background patients with no LC. The difference in the number of connections to patients with LC versus patients with no LC, called the “connection delta ratio” (CDR), was calculated for each health factor. Sorting the health factors by CDR in descending order provided a quantitative distribution of the health factors. Most of the top health factors with a CDR above a threshold were verified as LC risk factors or were correlated with LC in a literature review. This distribution laid the groundwork for grouping risk factors, selecting only 1 representative factor from each group for the ML model. For instance, pains at different body sites were combined into a single “pain” factor. Data for each variable group were also consolidated, considering the representative variable for the group as true if any of the variables in the group was true.

The following criteria were applied to select a small number of variables for the practical variable set: (1) ensuring that the number of essential variables remained fewer than 30 while achieving key prediction performance metrics (recall, precision, and accuracy) above 80%; (2) using consolidated variables based on the risk factor distribution wherever feasible; (3) minimizing the number of required laboratory tests; and (4) using imaging observations obtainable through simple chest radiographs. The rationale for these empirical criteria is to make the deployment and adoption of the model more practical in low-resource clinical settings, where data for only a small number of variables may be available. However, the LHS starting model should strike a balance between a minimal number of variables and acceptable performance metrics. We tested and compared feature selections using XGBoost. After determining a practical set, we ran RF, SVM, and KNN algorithms for comparison. All models were trained and evaluated using the default parameters of the classifiers. The XGBoost base model used the following default hyperparameters: `scale_pos_weight = 1`, `n_estimators = 500`, `max_depth = 6`, `eta = 0.3`, `gamma = 0`, `reg_lambda = 1.0`, `early_stopping_rounds = 5`, and `eval_metric = 'logloss'`.

Ethical Considerations

This retrospective study of EMR patient data received approval from the Institutional Review Board of Guilin Medical University Affiliated Hospital (number QTLL202139). Prior to data usage, our research team underwent training in patient data security and privacy policy of the hospital.

Results

Design of ML-LHS Unit for LC Screening

To improve patient inclusivity and adoption in LC screening, we designed a novel ML-enabled LHS unit for LC screening within a clinical research network (CRN). The CRN is led by a central hospital and participated by numerous clinics in surrounding communities and rural areas. The central hospital is tasked with developing an inclusive and practical LC risk prediction ML model to initialize the LHS unit and providing an AI tool online for clinic use. Primary physicians in these clinics are responsible for routinely using the AI tool to assess LC risk in all patient populations in the CRN. At-risk patients

are recommended for LC screening. The hospital also continuously updates models with new patient data, validates models, and deploys improved models for predictive services.

Inclusive LC Risk Prediction ML Models

A total of 2845 patients, comprising 1397 patients with LC and 1448 patients with no LC, were selected from the EMR of a Chinese hospital. The cohort consisted of 60.8% (1731/2845) men and 39.2% (1114/2845) women. Agewise, 19.6% (557/2845) patients were between 30 and 50 years of age, 58.1% (1654/2845) were between 50 and 70 years of age, and 22.0% (625/2845) were older than 70 years. Within the patient group with LC, 19.8% (277/2845) had a history of smoking, while 80.2% (1120/2845) did not. Since the data set includes a significant number of patients outside the typical LC-screening guideline-recommended demographic, which usually targets heavy smokers aged 50-80 years, the resulting LC risk prediction models were more inclusive, encompassing a broader patient population aged 30 years and older, regardless of smoking status.

To develop an LC risk prediction XGBoost model with default settings, we compared different sets of top-ranked health factors (including diseases, symptoms, medical histories, laboratory tests, observations, and other risk factors) from a list of more than 2000 factors, sorted by each factor’s prevalence in patients with LC. As the number of variables exceeded 200, key model performance metrics plateaued, reaching 0.85 for recall, 0.90 for precision, 0.88 for AUROC, and 0.88 for accuracy (Table 1 and Figure 2). Consequently, a set of 250 variables was selected as the inclusive variable set (denoted as “iv250”).

Using the iv250 set and default parameters, we compared XGBoost with other common algorithms such as RF, SVM, and KNN. Table 2 demonstrates that XGBoost and SVM achieved similarly high performance levels, with 0.86 for recall, 0.90 for precision, 0.89 for AUROC, and 0.89 for accuracy. The ROC curve and the reliability curve of the iv250 XGBoost model are shown in Figure 3.

Table 1. Performance metrics of the XGBoost lung cancer risk prediction models with different numbers of variables.

Metrics ^a	Number of variables									
	10	20	30	40	50	100	150	200	250	300
Recall	0.734	0.755	0.794	0.794	0.801	0.816	0.837	0.858	0.862	0.887
Precision	0.802	0.849	0.830	0.842	0.856	0.858	0.904	0.903	0.914	0.890
AUROC ^b	0.778	0.811	0.817	0.824	0.835	0.842	0.875	0.884	0.891	0.889
Accuracy	0.779	0.812	0.817	0.824	0.835	0.842	0.875	0.884	0.891	0.889

^aThe XGBoost machine learning base models were configured with default settings.

^bAUROC: area under the receiver operating characteristic curve.

Figure 2. Trends in performance metrics of XGBoost lung cancer risk prediction models with varying numbers of variables. Base models were trained using default settings. ROC-AUC: area under the receiver operating characteristic curve.

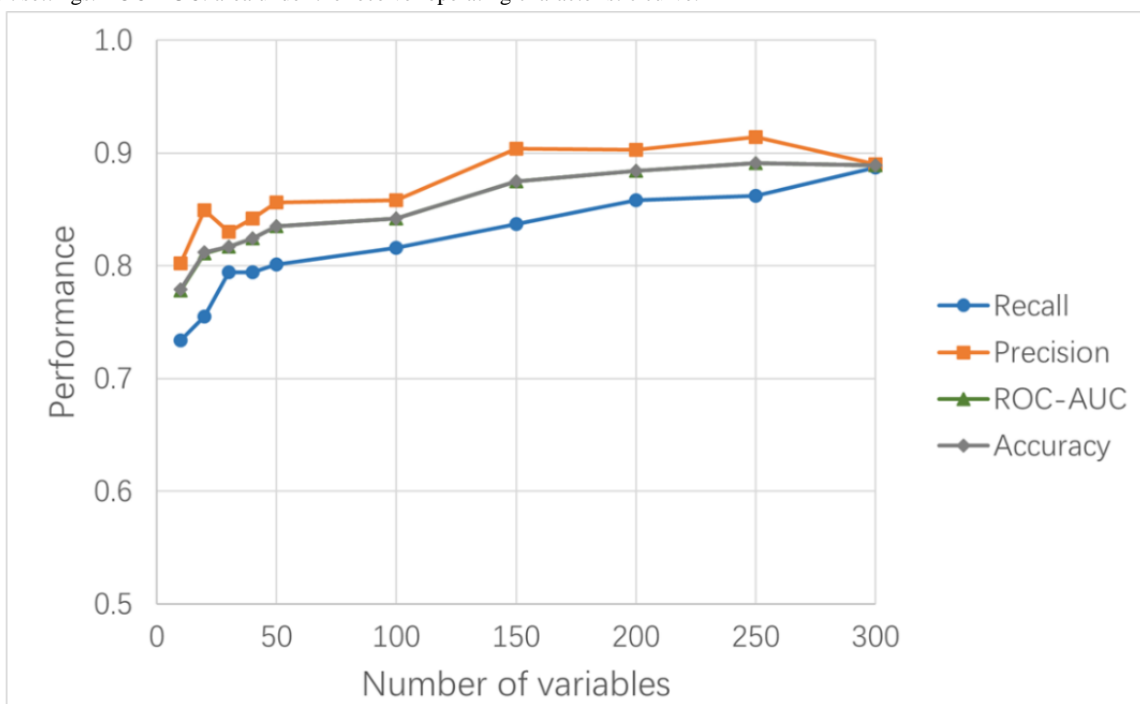


Table 2. Comparison of machine learning model performance using different algorithms for lung cancer risk prediction with default parameters^a.

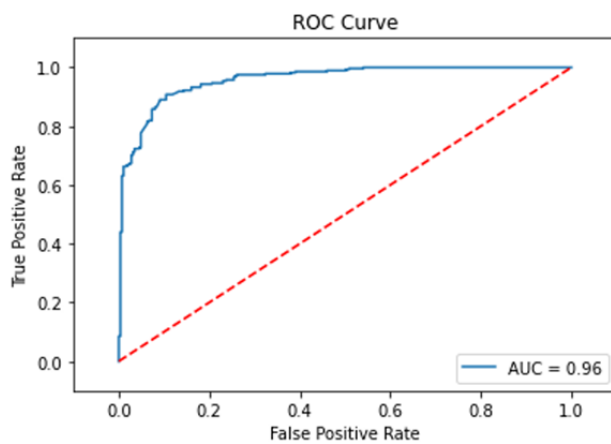
Algorithm	XGBoost	Random forest	Support vector machines	K-nearest neighbors
The inclusive 250-variable set (iv250)				
Recall	0.862	0.872	0.887	0.667
Precision	0.914	0.875	0.909	0.715
AUROC ^b	0.891	0.875	0.900	0.703
Accuracy	0.891	0.875	0.900	0.703
The inclusive and practical 29-variable set (pv29)				
Recall	0.805	0.816	0.748	0.649
Precision	0.825	0.830	0.858	0.832
AUROC	0.819	0.826	0.813	0.760
Accuracy	0.819	0.826	0.814	0.761

^aAll machine learning base models used default settings.

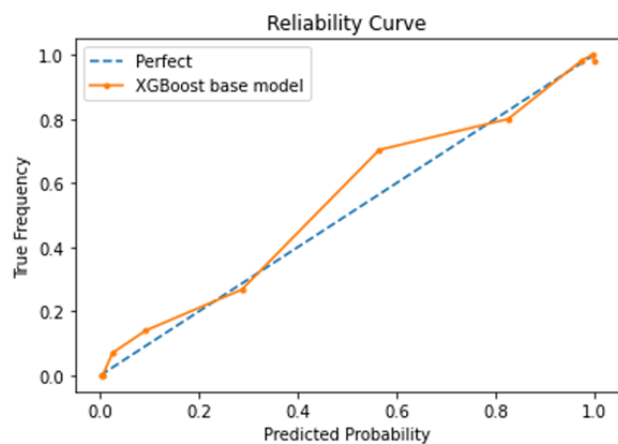
^bAUROC: area under the receiver operating characteristic curve.

Figure 3. ROC and reliability curves of XGBoost models for lung cancer risk prediction. Models were trained with the default settings. (A) ROC curve for the inclusive model using 250 variables (iv250). (B) Reliability curve for iv250. (C) ROC curve for the practical model using 29 variables (pv29). (D) Reliability curve for pv29. ROC: receiver operating characteristic.

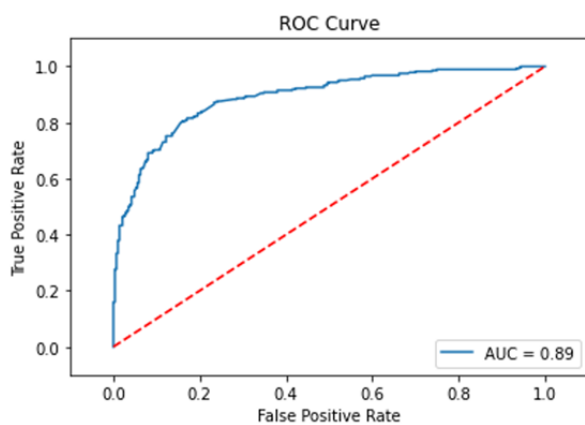
A.



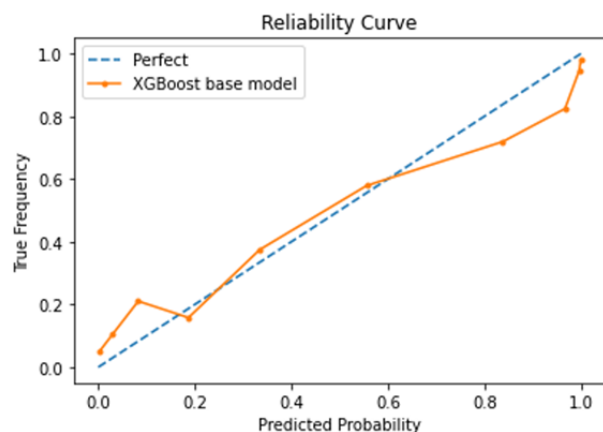
B.



C.



D.



Practical LC Risk Prediction ML Models

For practical application in clinics, the models underwent further refinement through feature engineering based on the quantitative

distribution of LC health factors. This refinement led to the development of a concise and practical set of 29 variables, termed “pv29.” Table 3 presents the details of the pv29 variables.

Table 3. List of the 29 variables used in the inclusive and practical machine learning models for lung cancer risk prediction.

Category	Local code	Health factor term
Disease	C-572430	Emphysema
Disease	C-654730	Lung inflammation
Disease	C-897420	Bronchitis
History	C-902187	Smoking history
Laboratory test	C-602395	Albumin/globulin ratio
Laboratory test	C-320164	Hematocrit
Laboratory test	C-952408	Non-small cell lung cancer-associated antigen
Laboratory test	C-023789	Carcinoembryonic antigen
Laboratory test	C-945807	Fibrinogen
Laboratory test	C-609483	Lymphocyte ratio
Laboratory test	C-346250	Platelet distribution width
Laboratory test	C-965710	Hemoglobin concentration
Laboratory test	C-546207	Globulin
Laboratory test	C-015328	Alkaline phosphatase
Laboratory test	C-963520	High-sensitivity C-reactive protein
Laboratory test	C-573086	Neuron-specific enolase
Laboratory test	C-284309	Carbohydrate antigen 153
Laboratory test	C-507246	Urine protein
Observation	C-598214	Lung nodules
Observation	C-825049	Pleural effusion
Observation	C-567942	Atelectasis
Risk factor	C-504168	Gender
Risk factor	C-928456	Age
Symptom	C-546879	Cough
Symptom	C-984012	Chest pain
Symptom	C-943817	Shortness of breath
Symptom	C-152064	Coughing up blood
Symptom	C-275809	Chest tightness
Symptom	C-549780	Pain

Table 2 compares the key performance metrics of the base models (XGBoost, RF, SVM, and KNN) using the pv29 variable set with default settings. The pv29 XGBoost and RF models demonstrated comparable performance, achieving 0.80 recall, 0.82 precision, 0.82 AUROC, and 0.82 accuracy. Figure 3 illustrates the ROC and reliability curves of the pv29 XGBoost model. Considering other requirements, including dealing with sparse data in EMRs and compute time, the pv29 XGBoost model was selected as the initial model for the LC risk prediction in initialization of the ML-LHS unit, aimed at the future

implementation of risk-based LC-screening recommendations in broader populations.

Discussion

Principal Findings

This study introduces a novel ML-LHS unit approach, aiming to offer sustainable and inclusive LC-screening solutions for all at-risk populations in both urban and rural areas within a CRN. To initiate this LC ML-LHS unit, we developed an inclusive and practical XGBoost model for LC risk prediction

using hospital EMR data. This enables risk-based LC screening in broader patient populations aged 30 years and older, regardless of smoking status. Using 29 variables, accessible even in low-resource clinics, the ML model achieved LC risk prediction with performance metrics of 0.80 recall, 0.82 precision, 0.82 AUROC, and 0.82 accuracy. Because most of the 29 variables were verified as risk factors or correlated factors for LC in literature, these model outputs are highly plausible. If an end user provides values for the 29 variables to the XGBoost model, the model will return a probability (0%-100%) of LC risk. More than 50% indicates a high risk of having LC, while below 50% indicates a low risk.

Future Direction: Implementing LC ML-LHS CRN

Considering the challenges in LC screening, such as low-screening adoption and inadequate coverage for nonsmokers and younger patients, exploring risk-based screening strategies is vital [11,33-35]. Following the present study, a future direction involves externally validating the LC risk prediction model. If validated, we plan to deploy the LC ML-LHS unit across a CRN, which will continuously monitor, rebuild the model, validate the new model, and deploy the improved model in so-called “LHS learning cycles.” Once operational, this innovative LHS unit could improve LC-screening rates and early detection in hospitals, community clinics, and rural areas.

Moreover, the ML-LHS CRN is well suited to screen for rare genetic mutations associated with LC, such as the ROS-1 mutation. If certain mutations are identified, personalized and precision medicine may be recommended by a doctor to the patient. Since the pv29 LC model does not contain the genetic mutations as variables, the LHS would need to integrate a large language model (LLM) into the prediction module for treatment prediction task. The top general-purpose LLMs, such as OpenAI's ChatGPT 4 and Google Gemini 1.5, have shown high accuracy in making medical predictions in our and many other studies without requiring structured data input [36,37]. Enhancing AI applicability through cooperation of structured data ML model and natural language LLMs presents an exciting future research direction.

Furthermore, screening is just the beginning of a patient's diagnostic journey in an equitable LHS. Future research should also investigate on how AI, particularly generative AI, and LHS can effectively follow up with high-risk patients, educate patients for shared decision-making, and remind patients to undergo diagnostic tests in time for early detection of LC. Simultaneously, LHS will coordinate primary care physicians and specialists to provide the appropriate diagnostics tests, such as image tests (computed tomography, positron emission tomography-computed tomography, and magnetic resonance imaging), pathology tests, and biopsies for final diagnosis. Future studies should also determine when to recommend molecular and genetic testing for achieving personalized and precision treatment.

Future Direction: Applying the ML-LHS Approach to Other Diseases

The vision of NAM's LHS emphasizes using RWD to generate real-world evidence. As EMRs are a primary source of RWD,

they can be used to develop inclusive and practical ML models for risk predictions of various diseases. Another promising future research direction is applying the ML-LHS unit approach proposed in this study to other preventable diseases and building LHS units in routine health care delivery, aimed at delivering more inclusive predictive screening in underserved populations.

We identify the biggest challenge of applying ML or AI in disease screening for all populations as the difficulty of deployment. ML models requiring a large number of variables may be deployed in hospitals, but they may not be usable in small clinics because the required data cannot be collected there. This study proposes a promising solution to this deployment problem: design a novel ML-enabled LHS unit and strike a balance of minimal variables and acceptable performance for the starting ML model of the LHS. Reducing the number of variables in a practical model usually reduces model performance compared with the inclusive mode. Setting 80% recall, precision, and accuracy as the acceptance bar, this study of the LC model and previous study of the nasopharyngeal cancer model demonstrated that it is possible to reduce the number of variables to below 30 [27].

For feature engineering, a common method is to use the feature importance list from the ML model. To meet the requirements of reducing variables to a minimal while keeping performance metrics above an acceptable level in starting up an ML-LHS unit, we have proposed an alternative approach that uses a quantitative distribution of health factors generated directly from EMR data by the patient graph CDR method in previous studies [26,27,32]. This study demonstrated again the effectiveness of the new feature selection approach of using health factor distribution from the CDR method in developing inclusive and practical ML models.

Limitations and Responsible AI

This study, however, has limitations. The EMR data presented issues with bias and missing data [38,39], which could potentially lead to biased models. For instance, smoking status and family history of LC were underreported in our data set. Significant efforts were made to understand and address these data biases, excluding variables where potential bias was identified. Despite these efforts, some biases may remain undetected and unmitigated. We also used algorithms such as XGBoost, known for effectively handling missing data. The lack of standardized structured data in EMRs made data collection labor-intensive. Reducing variables for practicality might risk overfitting in a small data set, though this issue should diminish as the ML-LHS unit continuously accumulates more data through its prediction service [40].

To further address these data bias issues as well as ML or AI application inequities, ML-LHS CRN will emphasize responsible AI development in future research [41]. First, CRN will strive to include more clinics from communities and rural areas surrounding the lead hospital, providing access to a broader population for AI-based LC screening. Second, the ML model will be frequently updated with new data from all patients, particularly including underserved populations, to continuously make the ML data set more representative and less biased. Third, a governance committee should be established

to review the development and use of the ML models to ensure high ethical standards, including protection of data safety and patient privacy, minimizing potential bias in data and algorithmic decision-making. Fourth, because mistakes or errors in AI prediction may cause harm or even deadly consequences, AI will be used only as a new information source for medical professionals or patients to make health care decisions.

Conclusions

This study devised an innovative ML-LHS unit for a CRN to sustainably offer inclusive LC screening to all at-risk

populations. For initializing such an ML-LHS unit serving community and rural clinics, we developed an inclusive and practical XGBoost model from hospital EMR data. Future deployment of the LC ML-LHS unit is expected to significantly improve LC-screening rates and early detection in broader populations, including those typically overlooked by existing LC-screening guidelines, such as nonsmokers and younger patients.

Acknowledgments

The authors would like to thank Mr Xiaowang Chen from the Department of Medical Information at Guilin Medical University Affiliated Hospital for his support with the EMR data server and privacy training. This work was supported by funding from the Guilin Municipal Science and Technology Bureau, China (grant 20190219-2), and the Sichuan Science and Technology Support Program, China (grant 2020YFQ0019).

Data Availability

The patient data sets used in the study are not available due to patient data privacy protection. Other data without privacy concern are available from the corresponding authors upon reasonable request.

Conflicts of Interest

None declared.

References

1. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global Cancer Statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin* 2021 May;71(3):209-249 [FREE Full text] [doi: [10.3322/caac.21660](https://doi.org/10.3322/caac.21660)] [Medline: [33538338](https://pubmed.ncbi.nlm.nih.gov/33538338/)]
2. Pinsky PF. Lung cancer screening with low-dose CT: a world-wide view. *Transl Lung Cancer Res* 2018 Jun;7(3):234-242 [FREE Full text] [doi: [10.21037/tlcr.2018.05.12](https://doi.org/10.21037/tlcr.2018.05.12)] [Medline: [30050762](https://pubmed.ncbi.nlm.nih.gov/30050762/)]
3. US Preventive Services Task Force, Krist AH, Davidson KW, Mangione CM, Barry MJ, Cabana M, et al. Screening for lung cancer: US Preventive Services Task Force Recommendation Statement. *JAMA* 2021 Mar 09;325(10):962-970. [doi: [10.1001/jama.2021.1117](https://doi.org/10.1001/jama.2021.1117)] [Medline: [33687470](https://pubmed.ncbi.nlm.nih.gov/33687470/)]
4. National Lung Screening Trial Research Team, Aberle DR, Adams AM, Berg CD, Black WC, Clapp JD, et al. Reduced lung-cancer mortality with low-dose computed tomographic screening. *N Engl J Med* 2011 Aug 04;365(5):395-409 [FREE Full text] [doi: [10.1056/NEJMoa1102873](https://doi.org/10.1056/NEJMoa1102873)] [Medline: [21714641](https://pubmed.ncbi.nlm.nih.gov/21714641/)]
5. Dubin S, Griffin D. Lung cancer in non-smokers. *Mo Med* 2020;117(4):375-379 [FREE Full text] [Medline: [32848276](https://pubmed.ncbi.nlm.nih.gov/32848276/)]
6. Thomas A, Chen Y, Yu T, Jakopovic M, Giaccone G. Trends and characteristics of young non-small cell lung cancer patients in the United States. *Front Oncol* 2015;5:113 [FREE Full text] [doi: [10.3389/fonc.2015.00113](https://doi.org/10.3389/fonc.2015.00113)] [Medline: [26075181](https://pubmed.ncbi.nlm.nih.gov/26075181/)]
7. Tammemägi MC, Katki HA, Hocking WG, Church TR, Caporaso N, Kvale PA, et al. Selection criteria for lung-cancer screening. *N Engl J Med* 2013 Feb 21;368(8):728-736 [FREE Full text] [doi: [10.1056/NEJMoa1211776](https://doi.org/10.1056/NEJMoa1211776)] [Medline: [23425165](https://pubmed.ncbi.nlm.nih.gov/23425165/)]
8. Tammemägi MC, Church TR, Hocking WG, Silvestri GA, Kvale PA, Riley TL, et al. Evaluation of the lung cancer risks at which to screen ever- and never-smokers: screening rules applied to the PLCO and NLST cohorts. *PLoS Med* 2014 Dec;11(12):e1001764 [FREE Full text] [doi: [10.1371/journal.pmed.1001764](https://doi.org/10.1371/journal.pmed.1001764)] [Medline: [25460915](https://pubmed.ncbi.nlm.nih.gov/25460915/)]
9. Yong PC, Sigel K, Rehmani S, Wisnivesky J, Kale MS. Lung cancer screening uptake in the United States. *Chest* 2020 Jan;157(1):236-238 [FREE Full text] [doi: [10.1016/j.chest.2019.08.2176](https://doi.org/10.1016/j.chest.2019.08.2176)] [Medline: [31916962](https://pubmed.ncbi.nlm.nih.gov/31916962/)]
10. Yang D, Liu Y, Bai C, Wang X, Powell CA. Epidemiology of lung cancer and lung cancer screening programs in China and the United States. *Cancer Lett* 2020;468:82-87. [doi: [10.1016/j.canlet.2019.10.009](https://doi.org/10.1016/j.canlet.2019.10.009)] [Medline: [31600530](https://pubmed.ncbi.nlm.nih.gov/31600530/)]
11. Wang Z, Wang Y, Huang Y, Xue F, Han W, Hu Y, et al. Challenges and research opportunities for lung cancer screening in China. *Cancer Commun (Lond)* 2018 Jun 07;38(1):34 [FREE Full text] [doi: [10.1186/s40880-018-0305-0](https://doi.org/10.1186/s40880-018-0305-0)] [Medline: [29880036](https://pubmed.ncbi.nlm.nih.gov/29880036/)]
12. Tammemägi MC. Selecting lung cancer screenees using risk prediction models-where do we go from here. *Transl Lung Cancer Res* 2018 Jun;7(3):243-253 [FREE Full text] [doi: [10.21037/tlcr.2018.06.03](https://doi.org/10.21037/tlcr.2018.06.03)] [Medline: [30050763](https://pubmed.ncbi.nlm.nih.gov/30050763/)]

13. Tammemagi MC, Schmidt H, Martel S, McWilliams A, Goffin JR, Johnston MR, PanCan Study Team. Participant selection for lung cancer screening by risk modelling (the Pan-Canadian Early Detection of Lung Cancer [PanCan] study): a single-arm, prospective study. *Lancet Oncol* 2017;18(11):1523-1531. [doi: [10.1016/S1470-2045\(17\)30597-1](https://doi.org/10.1016/S1470-2045(17)30597-1)] [Medline: [29055736](https://pubmed.ncbi.nlm.nih.gov/29055736/)]
14. Toumazis I, Bastani M, Han SS, Plevritis SK. Risk-based lung cancer screening: a systematic review. *Lung Cancer* 2020;147:154-186. [doi: [10.1016/j.lungcan.2020.07.007](https://doi.org/10.1016/j.lungcan.2020.07.007)] [Medline: [32721652](https://pubmed.ncbi.nlm.nih.gov/32721652/)]
15. Jonas DE, Reuland DS, Reddy SM, Nagle M, Clark SD, Weber RP, et al. Screening for lung cancer with low-dose computed tomography: updated evidence report and systematic review for the us preventive services task force. *JAMA* 2021;325(10):971-987. [doi: [10.1001/jama.2021.0377](https://doi.org/10.1001/jama.2021.0377)] [Medline: [33687468](https://pubmed.ncbi.nlm.nih.gov/33687468/)]
16. Pinsky P. Electronic health records and machine learning for early detection of lung cancer and other conditions: thinking about the path ahead. *Am J Respir Crit Care Med* 2021;204(4):389-390 [FREE Full text] [doi: [10.1164/rccm.202104-1009ED](https://doi.org/10.1164/rccm.202104-1009ED)] [Medline: [34097833](https://pubmed.ncbi.nlm.nih.gov/34097833/)]
17. Gould MK, Huang BZ, Tammemagi MC, Kinar Y, Shiff R. Machine learning for early lung cancer identification using routine clinical and laboratory data. *Am J Respir Crit Care Med* 2021;204(4):445-453. [doi: [10.1164/rccm.202007-2791OC](https://doi.org/10.1164/rccm.202007-2791OC)] [Medline: [33823116](https://pubmed.ncbi.nlm.nih.gov/33823116/)]
18. Wang X, Zhang Y, Hao S, Zheng L, Liao J, Ye C, et al. Prediction of the 1-year risk of incident lung cancer: prospective study using electronic health records from the state of Maine. *J Med Internet Res* 2019;21(5):e13260 [FREE Full text] [doi: [10.2196/13260](https://doi.org/10.2196/13260)] [Medline: [31099339](https://pubmed.ncbi.nlm.nih.gov/31099339/)]
19. Yeh MC, Wang Y, Yang H, Bai K, Wang H, Li YJ. Artificial intelligence-based prediction of lung cancer risk using nonimaging electronic medical records: deep learning approach. *J Med Internet Res* 2021;23(8):e26256 [FREE Full text] [doi: [10.2196/26256](https://doi.org/10.2196/26256)] [Medline: [34342588](https://pubmed.ncbi.nlm.nih.gov/34342588/)]
20. Institute of Medicine. *Best Care at Lower Cost: The Path to Continuously Learning Health Care in America*. Washington, DC: National Academies Press; 2013.
21. Institute of Medicine. *Digital Infrastructure for the Learning Health System: The Foundation for Continuous Improvement in Health and Health Care: Workshop Series Summary*. Washington, DC: National Academies Press; 2011.
22. Institute of Medicine. *The Learning Healthcare System: Workshop Summary*. Washington, DC: National Academies Press; 2007.
23. Simon GE, Platt R, Hernandez AF. Evidence from pragmatic trials during routine care—slouching toward a learning health system. *N Engl J Med* 2020;382(16):1488-1491. [doi: [10.1056/NEJMp1915448](https://doi.org/10.1056/NEJMp1915448)] [Medline: [32294344](https://pubmed.ncbi.nlm.nih.gov/32294344/)]
24. Institute of Medicine. *Large Simple Trials and Knowledge Generation in a Learning Health System: Workshop Summary*. Washington, DC: National Academies Press; 2013.
25. Chen A, Chen DO. Simulation of a machine learning enabled learning health system for risk prediction using synthetic patient data. *Sci Rep* 2022;12(1):17917 [FREE Full text] [doi: [10.1038/s41598-022-23011-4](https://doi.org/10.1038/s41598-022-23011-4)] [Medline: [36289292](https://pubmed.ncbi.nlm.nih.gov/36289292/)]
26. Chen A. A novel graph methodology for analyzing disease risk factor distribution using synthetic patient data. *Healthc Analytics* 2022;2:100084. [doi: [10.1016/j.health.2022.100084](https://doi.org/10.1016/j.health.2022.100084)]
27. Chen A, Lu R, Han R, Huang R, Qin G, Wen J, et al. Building practical risk prediction models for nasopharyngeal carcinoma screening with patient graph analysis and machine learning. *Cancer Epidemiol Biomarkers Prev* 2023;32(2):274-280. [doi: [10.1158/1055-9965.EPI-22-0792](https://doi.org/10.1158/1055-9965.EPI-22-0792)] [Medline: [36480263](https://pubmed.ncbi.nlm.nih.gov/36480263/)]
28. Chen T, Guestrin C. Xgboost: a scalable tree boosting system. 2016 Presented at: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 2016; San Francisco, CA p. 785-794.
29. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in Python. *JMLR* 2011;12(85):2825-2830.
30. Granger BE, Perez F. Jupyter: thinking and storytelling with code and data. *Comput Sci Eng* 2021;23(2):7-14. [doi: [10.1109/mcse.2021.3059263](https://doi.org/10.1109/mcse.2021.3059263)]
31. Liu Y, Chen PC, Krause J, Peng L. How to read articles that use machine learning: users' guides to the medical literature. *JAMA* 2019;322(18):1806-1816. [doi: [10.1001/jama.2019.16489](https://doi.org/10.1001/jama.2019.16489)] [Medline: [31714992](https://pubmed.ncbi.nlm.nih.gov/31714992/)]
32. Chen A, Huang R, Wu E, Han R, Wen J, Li Q, et al. The generation of a lung cancer health factor distribution using patient graphs constructed from electronic medical records: retrospective study. *J Med Internet Res* 2022;24(11):e40361 [FREE Full text] [doi: [10.2196/40361](https://doi.org/10.2196/40361)] [Medline: [36427233](https://pubmed.ncbi.nlm.nih.gov/36427233/)]
33. Sands J, Tammemägi MC, Couraud S, Baldwin DR, Borondy-Kitts A, Yankelevitz D, et al. Lung screening benefits and challenges: a review of the data and outline for implementation. *J Thorac Oncol* 2021;16(1):37-53 [FREE Full text] [doi: [10.1016/j.jtho.2020.10.127](https://doi.org/10.1016/j.jtho.2020.10.127)] [Medline: [33188913](https://pubmed.ncbi.nlm.nih.gov/33188913/)]
34. Tanner NT, Brasher PB, Wojciechowski B, Ward R, Slatore C, Gebregziabher M, et al. Screening adherence in the veterans administration lung cancer screening demonstration project. *Chest* 2020;158(4):1742-1752. [doi: [10.1016/j.chest.2020.04.063](https://doi.org/10.1016/j.chest.2020.04.063)] [Medline: [32439505](https://pubmed.ncbi.nlm.nih.gov/32439505/)]
35. Burnett-Hartman AN, Wiener RS. Lessons learned to promote lung cancer screening and preempt worsening lung cancer disparities. *Am J Respir Crit Care Med* 2020;201(8):892-893 [FREE Full text] [doi: [10.1164/rccm.201912-2398ED](https://doi.org/10.1164/rccm.201912-2398ED)] [Medline: [31905007](https://pubmed.ncbi.nlm.nih.gov/31905007/)]
36. Kanjee Z, Crowe B, Rodman A. Accuracy of a generative artificial intelligence model in a complex diagnostic challenge. *JAMA* 2023;330(1):78-80 [FREE Full text] [doi: [10.1001/jama.2023.8288](https://doi.org/10.1001/jama.2023.8288)] [Medline: [37318797](https://pubmed.ncbi.nlm.nih.gov/37318797/)]

37. Chen A, Chen DO, Tian L. Benchmarking the symptom-checking capabilities of ChatGPT for a broad range of diseases. *J Am Med Inform Assoc* 2023 Dec 18:ocad245. [doi: [10.1093/jamia/ocad245](https://doi.org/10.1093/jamia/ocad245)] [Medline: [38109889](https://pubmed.ncbi.nlm.nih.gov/38109889/)]
38. Kukhareva PV, Caverly TJ, Li H, Katki HA, Cheung LC, Reese TJ, et al. Inaccuracies in electronic health records smoking data and a potential approach to address resulting underestimation in determining lung cancer screening eligibility. *J Am Med Inform Assoc* 2022;29(5):779-788 [FREE Full text] [doi: [10.1093/jamia/ocac020](https://doi.org/10.1093/jamia/ocac020)] [Medline: [35167675](https://pubmed.ncbi.nlm.nih.gov/35167675/)]
39. Sauer CM, Chen L, Hyland SL, Girbes A, Elbers P, Celi LA. Leveraging electronic health records for data science: common pitfalls and how to avoid them. *Lancet Digit Health* 2022;4(12):e893-e898 [FREE Full text] [doi: [10.1016/S2589-7500\(22\)00154-6](https://doi.org/10.1016/S2589-7500(22)00154-6)] [Medline: [36154811](https://pubmed.ncbi.nlm.nih.gov/36154811/)]
40. Abraham E, Blanco C, Lee C, Christian J, Kass N, Larson E, et al. Generating knowledge from best care: advancing the continuously learning health system. In: *NAM Perspectives*. Washington, DC: National Academy of Medicine; 2016.
41. Goldberg CB, Adams L, Blumenthal D, Brennan PF, Brown N, Butte AJ, RAISE Consortium. To do no harm—and the most good—with AI in health care. *Nat Med* 2024;30(3):623-627. [doi: [10.1038/s41591-024-02853-7](https://doi.org/10.1038/s41591-024-02853-7)] [Medline: [38388841](https://pubmed.ncbi.nlm.nih.gov/38388841/)]

Abbreviations

AI: artificial intelligence
AUROC: area under the receiver operating characteristic curve
CDR: connection delta ratio
CRN: clinical research network
EMR: electronic medical record
KNN: K-nearest neighbors
LC: lung cancer
LHS: learning health system
LLM: large language model
ML: machine learning
ML-LHS: ML-enabled LHS
NAM: US National Academy of Medicine
PDJ: patient diagnosis journey
RF: random forest
ROC: Receiver operating characteristic curve
RWD: real-world data
SVM: support vector machines

Edited by K El Emam, B Malin; submitted 21.01.24; peer-reviewed by M Jani, B Green; comments to author 30.03.24; revised version received 02.04.24; accepted 01.05.24; published 11.09.24.

Please cite as:

Chen A, Wu E, Huang R, Shen B, Han R, Wen J, Zhang Z, Li Q

Development of Lung Cancer Risk Prediction Machine Learning Models for Equitable Learning Health System: Retrospective Study
JMIR AI 2024;3:e56590

URL: <https://ai.jmir.org/2024/1/e56590>

doi: [10.2196/56590](https://doi.org/10.2196/56590)

PMID: [39259582](https://pubmed.ncbi.nlm.nih.gov/39259582/)

©Anjun Chen, Erman Wu, Ran Huang, Bairong Shen, Ruobing Han, Jian Wen, Zhiyong Zhang, Qinghua Li. Originally published in JMIR AI (<https://ai.jmir.org>), 11.09.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR AI, is properly cited. The complete bibliographic information, a link to the original publication on <https://www.ai.jmir.org/>, as well as this copyright and license information must be included.

Original Paper

Mitigating Sociodemographic Bias in Opioid Use Disorder Prediction: Fairness-Aware Machine Learning Framework

Mohammad Yaseliani¹, MSc; Md Noor-E-Alam^{2,3}, PhD; Md Mahmudul Hasan^{1,4}, PhD

¹Department of Pharmaceutical Outcomes and Policy, College of Pharmacy, University of Florida, Gainesville, FL, United States

²Department of Mechanical and Industrial Engineering, Northeastern University, Boston, MA, United States

³The Institute for Experiential AI, Northeastern University, Boston, MA, United States

⁴Department of Information Systems and Operations Management, Warrington College of Business, University of Florida, Gainesville, FL, United States

Corresponding Author:

Md Mahmudul Hasan, PhD

Department of Pharmaceutical Outcomes and Policy, College of Pharmacy, University of Florida

Malachowsky Hall for Data Science & Information Technology, Suite 6300

1889 Museum Rd

Gainesville, FL, 32611

United States

Phone: 1 352 273 6276

Email: hasan.mdmahmudul@ufl.edu

Abstract

Background: Opioid use disorder (OUD) is a critical public health crisis in the United States, affecting >5.5 million Americans in 2021. Machine learning has been used to predict patient risk of incident OUD. However, little is known about the fairness and bias of these predictive models.

Objective: The aims of this study are two-fold: (1) to develop a machine learning bias mitigation algorithm for sociodemographic features and (2) to develop a fairness-aware weighted majority voting (WMV) classifier for OUD prediction.

Methods: We used the 2020 National Survey on Drug and Health data to develop a neural network (NN) model using stochastic gradient descent (SGD; NN-SGD) and an NN model using Adam (NN-Adam) optimizers and evaluated sociodemographic bias by comparing the area under the curve values. A bias mitigation algorithm, based on equality of odds, was implemented to minimize disparities in specificity and recall. Finally, a WMV classifier was developed for fairness-aware prediction of OUD. To further analyze bias detection and mitigation, we did a 1-N matching of OUD to non-OUD cases, controlling for socioeconomic variables, and evaluated the performance of the proposed bias mitigation algorithm and WMV classifier.

Results: Our bias mitigation algorithm substantially reduced bias with NN-SGD, by 21.66% for sex, 1.48% for race, and 21.04% for income, and with NN-Adam by 16.96% for sex, 8.87% for marital status, 8.45% for working condition, and 41.62% for race. The fairness-aware WMV classifier achieved a recall of 85.37% and 92.68% and an accuracy of 58.85% and 90.21% using NN-SGD and NN-Adam, respectively. The results after matching also indicated remarkable bias reduction with NN-SGD and NN-Adam, respectively, as follows: sex (0.14% vs 0.97%), marital status (12.95% vs 10.33%), working condition (14.79% vs 15.33%), race (60.13% vs 41.71%), and income (0.35% vs 2.21%). Moreover, the fairness-aware WMV classifier achieved high performance with a recall of 100% and 85.37% and an accuracy of 73.20% and 89.38% using NN-SGD and NN-Adam, respectively.

Conclusions: The application of the proposed bias mitigation algorithm shows promise in reducing sociodemographic bias, with the WMV classifier confirming bias reduction and high performance in OUD prediction.

(JMIR AI 2024;3:e55820) doi:[10.2196/55820](https://doi.org/10.2196/55820)

KEYWORDS

opioid use disorder; fairness and bias; bias mitigation; machine learning; majority voting

Introduction

Background

Opioid use disorder (OUD) and opioid overdose (OD) continue to remain a major public health crisis in the United States. OUD significantly contributes to overdoses, and >81,000 individuals lost their lives because of OD from April 2021 to April 2022 [1-3]. Meanwhile, the COVID-19 pandemic has worsened the ongoing OUD and OD epidemic [4]. In addition, the economic burden of OUD in the United States is overwhelming, with an estimated annual cost exceeding US \$786 billion in 2018 [1]. Therefore, it is critical to design interventions to facilitate an informed prescribing practice and monitoring of opioids to reduce the prevalence of OUD and subsequent drug overdose deaths.

Previous studies have used conventional regression-based methods to identify the significant predictors of OUD and OD [5-16]. More recently, machine learning (ML) methods have shown great potential in developing reliable predictive tools for identifying individuals at higher risk of OUD and OD [17]. ML methods can handle the complex nonlinear relationship among predictors and outcomes and perform well on imbalanced data [18]. Using different features as inputs, ML can predict the risk of developing OUD and OD with higher predictive accuracy [19,20]. Previous studies have developed random forests [16,21-24], decision trees [22,24], gradient boosting [16,23,25,26], neural networks (NNs) [5,16,27,28], and long short-term memory networks [24,28,29] to predict OUD and OD with impressive predictive performance.

Prior studies also reported that the risk of OUD or OD varies based on several individual-level protected sociodemographic features [30-32], potentially causing user-related bias. For instance, economically disadvantaged areas present higher levels of opioid use compared to other areas. Studies have highlighted significant sex differences in OUD in the United States, with women experiencing higher rates of prescription opioid use and faster progression to dependency compared to men [33,34]. In addition, opioid prescribing is 2 times more likely for White individuals than Black individuals in the United States [35,36], attributed to physicians' practice biases [37]. Therefore, the real-world data describing OUD patterns often include biases caused by users. These biases, alongside potential sampling or algorithmic biases, could cause unfair and biased outcomes, leading to suboptimal model performance and inequalities in patient care [38].

Objectives

In this study, we aim to address the limitations of prior studies, including the lack of attention to (1) detecting and analyzing the fairness and bias in the ML or deep learning (DL) models for predicting OUD and (2) proposing methods to mitigate bias for different protected attributes. Using the data provided by the National Survey on Drug Use and Health (NSDUH) for 2020, we developed an NN model to detect the bias for different sociodemographic features [39,40]. We then propose an algorithm based on equality of odds (EO) to mitigate the bias while ensuring reasonable predictive performance (ie, accuracy and recall). Finally, we create a fairness-aware weighted

majority voting (WMV) classifier that considers the predicted classes using the optimal thresholds for different sociodemographic features and outputs the most frequent class. To show the effectiveness of the proposed methods, we also evaluate their performance by developing several ML algorithms, including logistic regression (LR), linear support vector machine (SVM), and SVM-radial basis function (SVM-RBF).

Methods

Data and Sample

We used the 2020 NSDUH survey data that was conducted using an independent multistage area probability design [41]. The study sample included a community-based noninstitutionalized population aged ≥ 12 years, with information on clinical characteristics, sociodemographic factors, and substance use. The final data included 32,893 individuals with 2892 variables for public use.

Features and Outcome Variable

We selected the features based on the prior research identifying predictors of OUD [42-52]. The sociodemographic features included sex, marital status, working condition (whether someone works ≥ 35 h/wk), race (Black, White, and other racial groups), and income (<US \$20,000 per year and other groups). We also included a history of using different types of prescription opioids [42] (eg, oxycodone, oxymorphone, hydrocodone, hydromorphone, fentanyl, morphine, codeine, methadone, tramadol, and buprenorphine), use of heroin, history of receiving alcohol or drug treatment, diabetes [43], chronic bronchitis [44], cirrhosis of the liver [45], hepatitis B or C [46,47], kidney disease [48], asthma [49], AIDS [50], cancer [51], depression [52], and BMI [53,54]. A total number of 26 features were included in the proposed ML and DL models. After one-hot encoding, there were a total of 44 features, including BMI. These features are presented in [Multimedia Appendix 1](#).

As an outcome variable, we used whether an individual has developed OUD, which is defined as the dependence, misuse, and abuse of opioids [55]. To train the classifiers, we used stratified 80:20 train-test splitting. Of the 26,314 individuals, the training set included 26,148 (99.37%) and 166 (0.63%) individuals belonging to non-OUD and OUD classes, respectively. Moreover, of the 6579 individuals, there were 6538 (99.38%) non-OUD and 41 (0.62%) OUD individuals in the test set.

Classifiers

To perform this prediction task, we primarily developed and evaluated 3 well-known ML or DL models: LR [56], SVM [57], and NN [58]. We first designed an NN model consisting of 4 layers, in which the first layer takes 44 features as inputs, the 2 hidden layers include 1000 neurons, and the last layer consists of 1 neuron for making predictions. We implemented this NN model using the TensorFlow library in Python created by Abadi et al [59] using minibatch stochastic gradient descent (SGD) and the Adam optimizer, using the default learning rate of 0.001. Minibatch SGD was chosen to optimize the convergence speed

and computational efficiency, as it balances the stability of full-batch gradient descent and the high variance of SGD, enabling more robust learning [60]. The Adam optimizer was used for its adaptive learning rate properties, which adjust the learning rate for each parameter individually. This optimizer combines the benefits of AdaGrad and RMSProp, making it effective for handling sparse gradients and nonstationary objectives [61]. A batch size of 64 was used to leverage parallelism on modern hardware, thus reducing computational time and improving scalability. The NNs were trained for 20 epochs using the binary cross-entropy loss function with balanced class weighting to address any class imbalance in the data set, ensuring that the model is not biased toward the majority class and performs well on minority class instances.

As noted before, we also trained ML classifiers to test the performance of the proposed methods. To train the LR classifier, we fitted intercept and used the balanced class weighting to prevent overfitting or underfitting. To train the SVM classifiers, we used a C value of 1 with balanced class weighting. The default γ value for the RBF kernel in the SVM-RBF classifier was calculated according to equation 1:



(1)

Notably, we used recall, specificity, and accuracy to assess the performance of all these models.

Bias Detection Based on the Area Under the Curve Values

We largely followed the study by Fletcher et al [39] to detect both the algorithmic and sampling bias for each of the sociodemographic features (ie, sex, marital status, working condition, race, and income) based on the area under the curve (AUC) value. We computed the AUC values for each category of a sociodemographic feature, where any difference in the AUC values identified the presence of algorithmic bias. This type of bias is the result of internal model specifications and features [39].

Sampling bias detection and mitigation are critical in ensuring the fairness and effectiveness of ML models, especially in the context of global health. As discussed in the study by Fletcher et al [39], sampling bias arises when the data used to train an ML model does not adequately represent the actual proportions found in the real world. This can lead to models that perform poorly on minority groups and introduce unfairness into the decision-making process. The methodology for detecting sampling bias involves creating homogenous test groups for each demographic category and comparing the model's performance. By examining the AUC and the variability of model performance across these groups, the biases caused by sampling can be identified. For instance, if a model trained predominantly on data from one demographic group consistently underperforms when applied to another group, this indicates the presence of sampling bias. Such disparities highlight the model's inability to generalize well across diverse populations, which is a fundamental flaw in its design. In our study, we

systematically tested for sampling bias by examining the model's performance across different demographic groups.

To detect the sampling bias, we first created training sets consisting of different compositions of individuals belonging to each sociodemographic group. We then computed and plotted the models' AUC value for each category based on a fixed-size test set. A significant fluctuation in the models' AUC values with respect to the change in the structure of the training sets indicated the presence of the sampling bias for a sociodemographic feature [39].

In this study, we analyzed 5 sociodemographic features, including sex, marital status, working condition, race, and income, to detect the algorithmic bias. The categories included male individuals and female individuals, those who have never been married and other groups related to marital status, those who work ≥ 35 hours per week and other groups associated with working condition, Black and White race, and those who have an income of $< US \$20,000$ per year and others. To detect the sampling bias, we created a test set of fixed size for each sociodemographic feature: sex: 5984 (2992 male individuals and 2992 female individuals), marital status: 5648 (2824 from *never been married* and 2824 from other groups), working condition: 5296 (2648 from *working ≥ 35 hours* and 2648 from other groups), race: 1280 (640 Black race and 640 from White race), and income: 2000 (1000 from *income $< US \$20,000$* and 1000 from other groups). These test sets were created based on the main test set used during model development (6579/32,893, 20% of the whole data).

To further enhance the validity of the proposed predictive model, we implemented a detailed 1-N matching process, in which we controlled for a variety of socioeconomic variables, including BMI, sex, marital status, working condition, race, and income, to achieve a balanced comparison between OUD and non-OUD cases. To ensure robust matching, we used a 1-N matching strategy with $n=158$, meaning each OUD case was matched with 158 non-OUD cases. This number was calculated based on the proportion of OUD-negative to OUD-positive cases in the 2020 NSDUH data used in this study. Following 1-N matching, the training set included 26,164 non-OUD and 166 OUD individuals. Similarly, 6542 individuals belonged to the non-OUD class in the test set and 41 were in the OUD class. Accordingly, we implemented the same sampling and algorithmic bias detection approaches on the matched data to comprehensively analyze the existence of bias in the predictive models. This extensive matching allowed us to control potential confounders and provide a more accurate data set for our analysis.

The Proposed Method for Bias Mitigation

To propose a method for bias mitigation, we considered a fairness definition called EO [62], which is measured based on the difference between the specificity and recall or sensitivity values. This approach is grounded in the study by Hardt et al [62], in which their framework introduces a robust criterion for measuring and removing discrimination based on protected attributes, emphasizing that fairness can be optimized post hoc through threshold adjustments. They proposed that any learned predictor can be adjusted to achieve nondiscrimination by

modifying its thresholds, ensuring that the predictor's true positive rates (recall) and false positive rates (1-specificity) are independent of the protected attribute, thereby satisfying the EO criterion. This framework suggests that EO can be achieved without altering the underlying complex training pipeline of the predictor. Instead, a simple postprocessing step is sufficient, which is both practical and efficient. This method is robust to changes in class distribution and ensures that the model remains fair by balancing the sensitivity and specificity across different groups. By focusing on minimizing the difference between recall and specificity, we ensure that our model does not favor one group over another, thereby maintaining fairness and robustness in our predictions. The postprocessing adjustment of thresholds allows us to achieve these fairness criteria without compromising the utility of the model, providing a balanced approach to bias mitigation. Equation 2 demonstrates the formula of EO,



Textbox 1. Algorithm 1.

Input: A trained machine learning (or deep learning)-based model for the prediction of opioid use disorder.

Output: An optimal threshold value.

Begin

1 Z_Values = []

2 th_Values = []

3 th ← 0

4 i ← 0

5 While th ≤ 100 do

6 Calculate the overall recall and accuracy of the model.

7 If recall ≥ 0.7 and accuracy ≥ 0.5 then

8 Calculate x_1 and x_2 (specificity values), and y_1 and y_2 (recall values)

9 Calculate

10 Append (Z_Values, Z)

11 Append (th_Values, th)

12. th ← th + 0.1

13 For $i = 1$ to 1001 do

14 if Z_Values[i] = Min(Z_Values) then

15 Best_Threshold th_Values[i]

End

The Proposed WMV Classifier

Algorithm 1 (Textbox 1) will output different classification thresholds for each sociodemographic feature. Therefore, we may achieve different outputs or classes for a given individual using those thresholds. However, in most cases, we need to have a predictive model that takes into account the bias-related issues for multiple sociodemographic features and predicts OUD for a given individual such that the bias is mitigated to the greatest extent. In this regard, we propose a WMV classifier,

(2)

where G denotes the group being analyzed and y represents the output class. This is equivalent to balancing the recall and specificity values of both groups and considering them equal.

To achieve the EO, we change the classification threshold [62] so that the difference between the recall and specificity is minimized. To address the decreased performance measures (ie, recall and accuracy) because of the threshold moving, we define minimum values for recall and accuracy to ensure that they are above a certain value while changing the classification threshold (70% for recall and 50% for accuracy). We tested the threshold values in the range (0, 100) and identified the optimal one where the recall and accuracy constraints are satisfied and the difference between specificity and recall is minimum. Algorithm 1 (Textbox 1) presents the details of the proposed bias mitigation method. The input is a trained ML or DL model for OUD prediction, and the output is an optimal threshold value based on which the classifications are performed.

which yields a class based on the classifications by each threshold according to equation 3,

$$P = w_1 \times pc_1 + w_2 \times pc_2 + \dots + w_n \times pc_n \quad (3)$$

where P is representative of the final probability, pc_i shows the class predicted (ie, 0 and 1) using the threshold for a sociodemographic feature I , and w_i shows the weight assigned to feature i . To assign a weight to each feature, we calculate the proportion of differences between recall and specificity to gain

a value in the range [0,1] and acquire the final probability using equation 3.

Ethical Considerations

This study did not require approval from an institutional ethics board or review committee as it did not involve any human participants or identifiable personal data. The study used publicly available, anonymized data sets that are confidential and only used for statistical purposes per federal law.

Results

Overview

We considered the NN models trained with SGD and Adam optimizers (NN model using SGD [NN-SGD] and NN model using Adam [NN-Adam]) as the main classifiers and reported results that we obtained for bias detection and mitigation for predicting OUD. We also described the results for 3 ML classifiers (ie, LR, linear SVM, and SVM-RBF) in [Multimedia Appendix 2](#).

Individual Characteristics

The individuals in the training (26,314/32,893, 80%) and test (6579/32,893, 20%) samples had similar sociodemographic and

clinical features ([Multimedia Appendix 1](#)). The mean BMI of individuals was 25.58 (95% CI 25.49-25.68; SD 6.68), 54% (17,763/32,893) of individuals were female, and 46% (15,130/32,893) had developed OUD. While the least used opioid was oxycodone (95/32,893, 0.29%), the most commonly used opioids were hydrocodone (3495/32,893, 10.63%), followed by oxycodone (2147/32,893, 6.53%) and codeine (2014/32,893, 6.12%). In addition, approximately 4.87% (1602/32,893) of individuals had the experience of receiving drug treatments. In total, 28.82% (9482/32,893) of individuals had a history of depression, followed by asthma (3934/32,893, approximately 11.96%) and diabetes (1848/32,893, 5.62%). In addition, >36.93% (12,146/32,893) of individuals were married, and approximately 40.77% (13,409/32,893) of individuals worked for ≥ 35 hours per week. In addition, approximately 9.19% (3025/32,893) of individuals were Black, 64.94% (21,362/32,893) were White, and the rest belonged to other races (8506/32,893, 25.86%). Furthermore, while approximately 14.95% (4917/32,893) of individuals had an income of <US \$20,000 per year, almost 85.05% (27,976/32,893) earned >US \$20,000 yearly. [Table 1](#) summarizes the sociodemographic features used in the study.

Table 1. The details of sociodemographic variables in the study (N=32,893)^a.

Sociodemographic variables	Total	Non-OUD ^b (n=32,686, 99.37%)	OUD (n=207, 0.63%)
Sex, n (%)			
Male	15,130 (46)	15,024 (45.96)	106 (51.21)
Female	17,763 (54)	17,662 (54.04)	101 (48.79)
Marital status, n (%)			
Married	12,146 (36.93)	12,101 (37.02)	45 (21.74)
Widowed	661 (2.01)	659 (2.02)	2 (1)
Divorced or separated	2613 (7.94)	2569 (7.86)	44 (21.26)
Never been married	14,452 (43.94)	14,346 (43.89)	106 (51.21)
Working condition, n (%)			
>35 hours per week	13,409 (40.77)	13,356 (40.86)	53 (25.6)
<35 hours per week	4644 (14.12)	4619 (14.13)	25 (12.08)
Race, n (%)			
White	21,362 (64.94)	21,213 (64.9)	149 (71.98)
Black	3025 (9.2)	3008 (9.2)	17 (8.21)
Other racial groups	8506 (25.86)	8465 (25.9)	41 (19.81)
Income (US \$), n (%)			
<\$20,000	4917 (14.95)	4851 (14.84)	66 (31.88)
Other income groups	27,976 (85.05)	27,835 (85.16)	141 (68.12)
BMI (kg/m ²), mean (SD)	25.59 (8.82)	25.59 (8.82)	25.16 (9.83)

^aFor marital status and working conditions, we excluded those not answering the questionnaire or those who legitimately skipped the question, as described in the National Survey on Drug Use and Health data dictionary.

^bOUD: opioid use disorder.

Performance of the Classifiers

Figure 1 shows the receiver operating characteristic (ROC) and precision-recall (PR) curves of the trained ML or DL classifiers for OUD prediction. As shown in Figure 1, while the SVM-RBF classifier has the highest ROC/AUC (AUC 97.13%, 95% CI 93.53%-100%), NN-Adam performs best in terms of PR/AUC (AUC 38.29%, 95% CI 37.11%-39.46%). Moreover, the NN-Adam outperforms the NN-SGD (ROC/AUC 85.66%, 95% CI 78.36%-92.95%; ROC/PR 7.10%, 95% CI 6.48%-7.72%) and NN-Adam (ROC/AUC 96.57%, 95% CI 92.67%-100.00%; ROC/PR 38.29%, 95% CI 37.11%-39.46%).

Figure 2 shows the ROC and PR curves of the ML or DL classifiers for OUD prediction after matching. As shown in Figure 2, the SVM-RBF classifier has the highest ROC/AUC (AUC 97.13%, 95% CI 93.53%-100%), while LR has the highest PR/AUC (AUC 28.70%, 95% CI 27.60%-29.79%). Moreover, while NN-SGD outperforms NN-Adam in terms of ROC/AUC, NN-Adam has a higher PR/AUC (NN-SGD: ROC/AUC 94.95%, 95% CI 90.27%-99.63%; ROC/PR 18.13%, 95% CI:

17.20%-19.06%; NN-Adam: ROC/AUC 92.47%, 95% CI 86.86%-98.07%; ROC/PR 23.25%, 95% CI 22.23%-24.27%).

Figure S1 in Multimedia Appendix 2 reports the confusion matrices of NN models. While the NN-SGD correctly classifies only 22% (9/41) of the individuals who have developed OUD, the NN-Adam correctly classifies 71% (29/41) of individuals. Moreover, the NN-SGD misclassifies 0.54% (35/6538) of the individuals who have not developed OUD, whereas the NN-Adam misclassifies 2.68% (175/6538) of individuals. Overall, the NN-SGD and NN-Adam achieve a recall of 21.95% and 70.73%, a specificity of 99.46% and 97.32%, and an accuracy of 98.98% and 97.16%, respectively.

The confusion matrices of the ML classifiers, including LR, linear SVM, and SVM-RBF, are presented in Figure S2 in Multimedia Appendix 2. All these classifiers have an accuracy/specificity of >92% and an AUC of >96%. Moreover, the linear SVM classifier has the highest recall of 82.93%, followed by LR and SVM-RBF with 80.49% and 26.83%, respectively.

Figure 1. The receiver operating characteristic and precision-recall curves of the classifiers. (A) The receiver operating characteristics curve of the classifiers, (B) precision-recall curve of the classifiers. AUC: area under the curve; LR: logistic regression; NN: neural network; RBF: radial basis function; SGD: stochastic gradient descent; SVM: support vector machine.

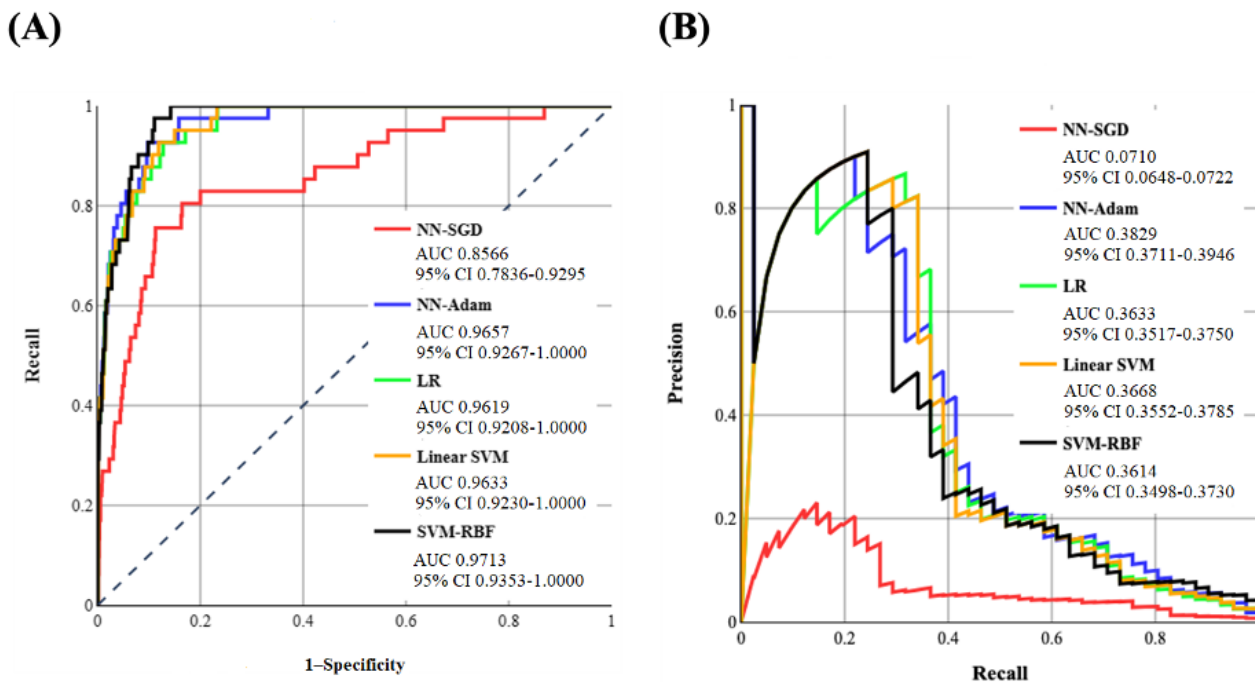
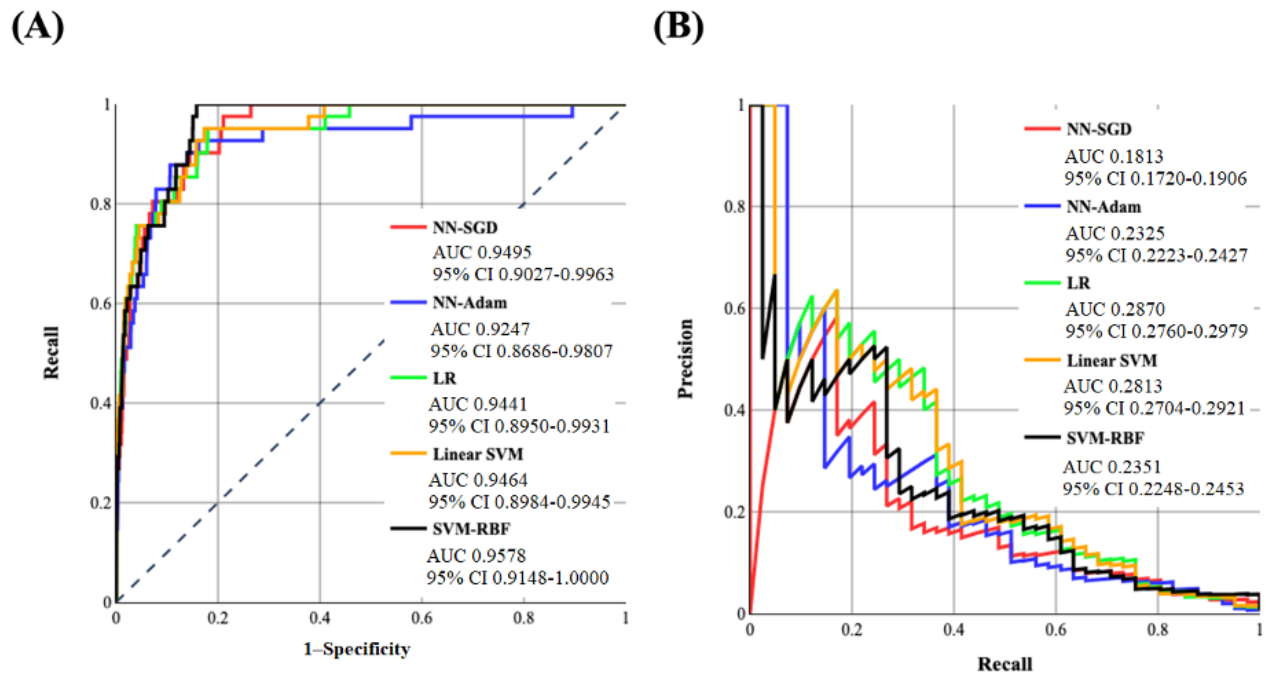


Figure 2. The receiver operating characteristic and precision-recall curves of the classifiers after matching. (A) The receiver operating characteristics curve of the classifiers, (B) precision-recall curve of the classifiers. AUC: area under the curve; LR: logistic regression; NN: neural network; RBF: radial basis function; SGD: stochastic gradient descent; SVM: support vector machine.



1-N Matching

The standardized mean difference and variance ratio before and after matching for different sociodemographic groups are presented in [Table 2](#).

Table 2. The details of 1-N matching (N=158).

Status	SMD ^a	Variance ratio
Race		
White		
Before matching	0.1526	0.8896
After matching	0.0188	0.9869
Black		
Before matching	-0.0351	0.9065
After matching	-0.0385	0.8979
Other		
Before matching	-0.1453	0.8317
After matching	0.0060	1.0141
Income		
<US \$20,000 per year		
Before matching	0.4106	1.7267
After matching	-0.0178	0.9913
Sex		
Male		
Before matching	0.1049	1.0108
After matching	0.0069	1.0045
Marital status		
Married		
Before matching	-0.3400	0.7332
After matching	0.0094	1.0180
Widowed		
Before matching	-0.0866	0.4867
After matching	-0.0019	0.9863
Divorced or separated		
Before matching	0.3862	2.3224
After matching	0.0324	1.0534
Never married		
Before matching	0.1467	1.0195
After matching	-0.0077	1.0053
Working condition		
Working ≥35 hours per week		
Before matching	-0.3279	0.7921
After matching	0.0391	1.0520
Other groups of working condition		
Before matching	-0.0608	0.8793
After matching	0.0069	1.0212
BMI		
Before matching	-0.0464	1.2435
After matching	0.0169	1.0244

^aSMD: standardized mean difference.

Bias Detection (Algorithmic Bias)

Figures 3 and 4 demonstrate the ROC curves for these abovementioned sociodemographic groups after training the NN-SGD and NN-Adam, respectively. Tables 3 and 4 show the performance metrics using the default threshold (50%) with *P* values for the difference between the specificity and recall.

According to Table 3, there is a high difference between the AUC values for both groups related to each of the 5 sociodemographic features, and the algorithmic bias was present in the NN-SGD [39]. Furthermore, the *P* values with 95% CI indicate that there is a statistically significant difference between specificity and recall values using various thresholds in the range (0, 100).

According to Table 4, the difference between AUC values for sociodemographic groups is high for all 5 sociodemographic features. Moreover, the accuracy and specificity values are notably high for all groups, and recall values are >57% (except for the Black race, which is 33.33%), which shows the high performance of the model in correctly identifying those with OUD. Similar to the NN-SGD, the *P* values are quite significant, and algorithmic bias is present in the NN-Adam.

The results of detecting algorithmic bias using LR, linear SVM, and SVM-RBF classifiers are presented in Figures S3-S5 in

Multimedia Appendix 2, respectively. Tables S1-S3 in Multimedia Appendix 2 also show the performance of these classifiers for sociodemographic features. All the classifiers indicate algorithmic bias. Moreover, while the SVM-RBF classifier indicates the highest bias for race, LR and linear SVM classifiers show a higher bias for sex and marital status.

Figures 5 and 6 demonstrate the ROC curves for sociodemographic groups after doing 1-N matching for the NN-SGD and NN-Adam, respectively. Tables 5 and 6 show the performance metrics using the default threshold (50%) with *P* values for the difference between the specificity and recall.

According to Table 5, there is a high difference between the AUC values for each sociodemographic feature, highlighting the existence of algorithmic bias in the NN-SGD [39]. Furthermore, there is a statistically significant difference between specificity and recall values according to *P* values.

According to Table 6, the difference between AUC values for sociodemographic groups is high for all 5 sociodemographic features. Moreover, the accuracy and specificity values are >53% for all groups, except for the Black race, demonstrating the high performance of the model in correctly identifying those with OUD. Furthermore, the *P* values are statistically significant, indicating the existence of algorithmic bias in the NN-Adam.

Figure 3. The receiver operating characteristic (ROC) curves for various groups related to sociodemographic features using the neural network model using stochastic gradient descent (with area under the curve [AUC] and 95% CI values). Values were calculated based on the test sample (6579 individuals: 41 developed opioid use disorder [OUD] and 6538 did not develop OUD). (A) ROC curve for sex, (B) ROC curve for marital status, (C) ROC curve for working conditions, (D) ROC curve for race, and (E) ROC curve for income.

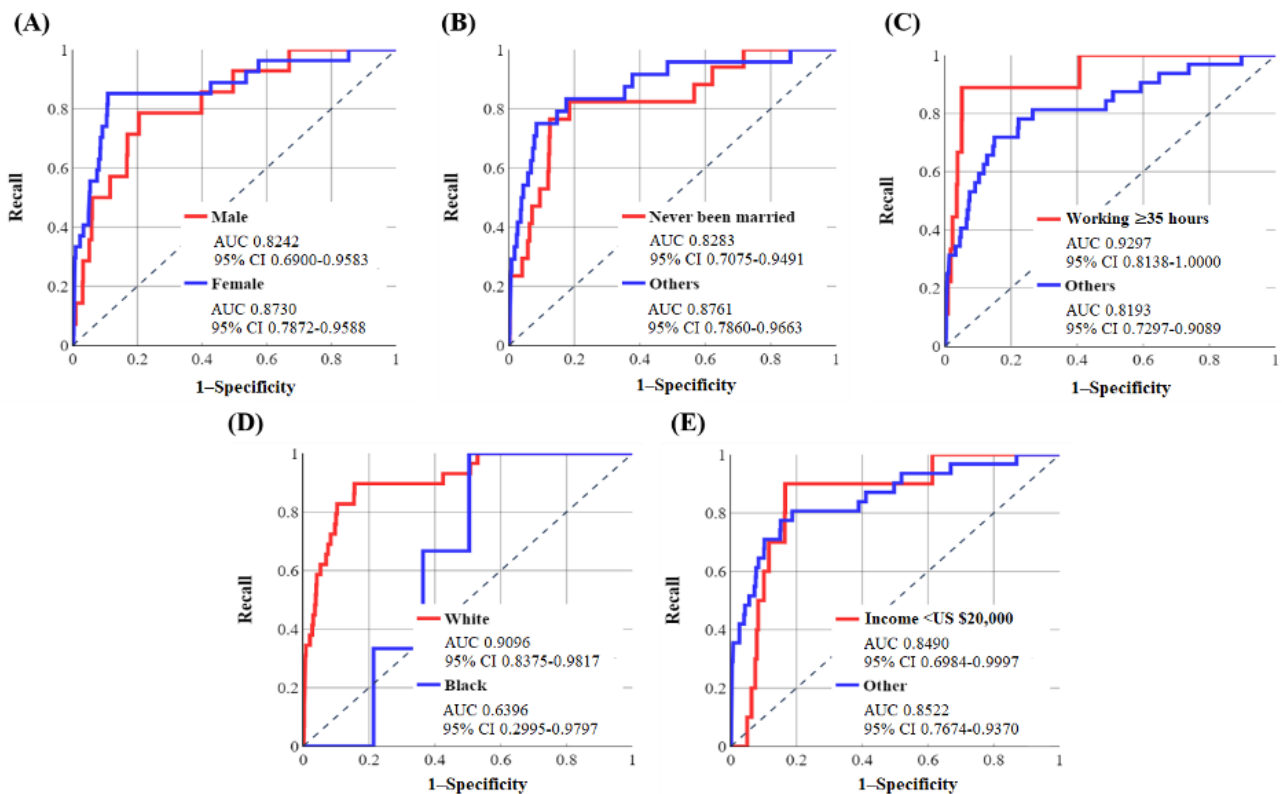


Figure 4. The receiver operating characteristic (ROC) curves for various groups related to sociodemographic features using the neural network model using Adam (with area under the curve [AUC] and 95% CI values). Values were calculated based on the test sample (6579 individuals: 41 developed opioid use disorder [OUD] and 6538 did not develop OUD). (A) ROC curve for sex, (B) ROC curve for marital status, (C) ROC curve for working condition, (D) ROC curve for race, and (E) ROC curve for income.

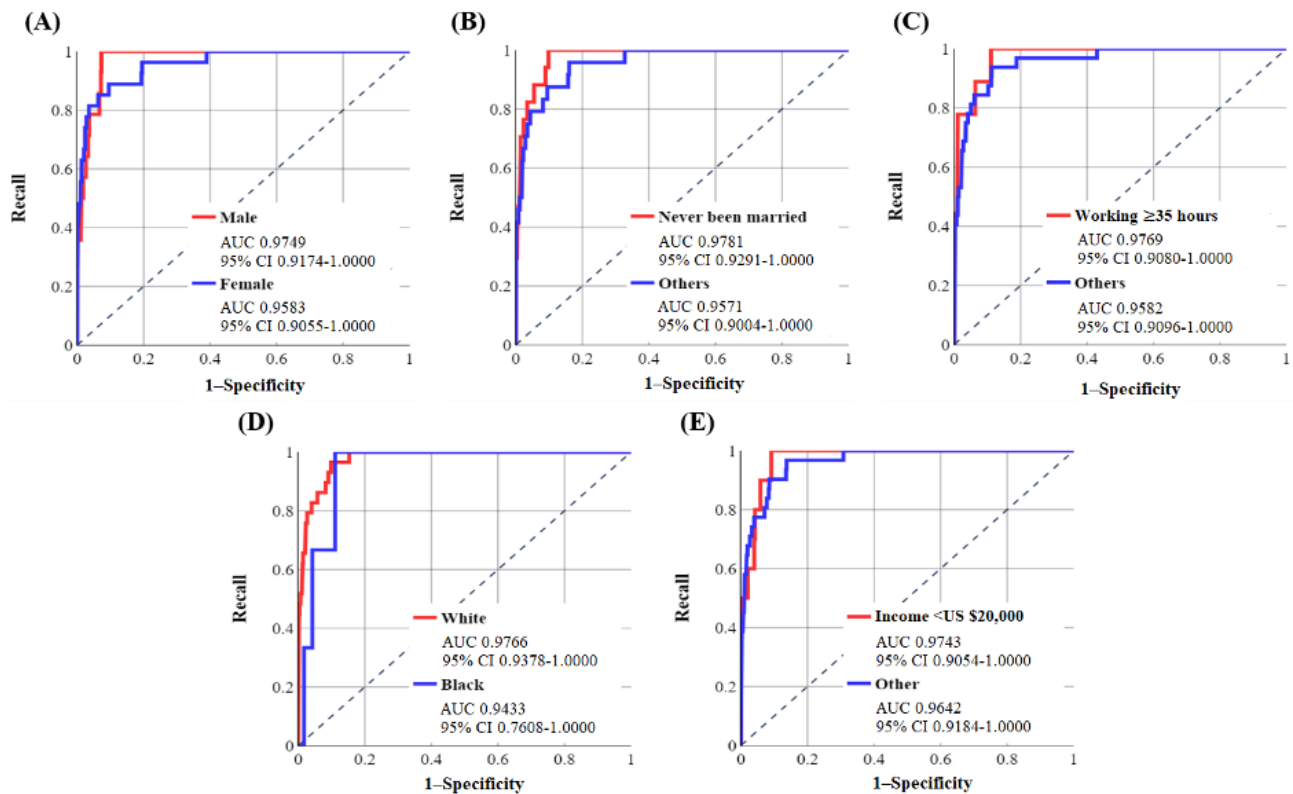


Table 3. The performance metrics of the neural network model using stochastic gradient descent using the default threshold (50%).

Sociodemographic groups	Recall	Specificity	Accuracy	AUC ^a	Difference ^b	<i>P</i> value
Sex					22.74	<.001
Male	7.14	99.60	99.17	82.42		
Female	29.63	99.35	98.83	87.30		
Marital status					3.07	<.001
Never been married	23.53	99.25	98.80	82.83		
Other groups of marital status	20.83	99.62	99.12	87.61		
Working condition					14.40	<.001
Working ≥ 35 hours	11.11	99.77	99.47	92.97		
Other groups of working condition	25.00	99.26	98.65	81.93		
Race					28.08	<.001
White	27.59	99.55	99.07	99.07		
Black	0.00	99.06	98.60	98.60		
Income					29.93	<.001
An income of <US \$20,000	0.00	98.70	97.72	84.90		
Other groups of income	29.03	99.60	99.21	85.22		

^aAUC: area under the curve.

^bDifference between recall and specificity values.

Table 4. The performance metrics of the neural network model using Adam using the default threshold (50%).

Sociodemographic groups	Recall	Specificity	Accuracy	AUC ^a	Difference ^b	P value
Sex					21.31	<.001
Male	57.14	97.69	97.50	97.49		
Female	77.78	97.02	96.87	95.83		
Marital status					10.40	<.001
Never been married	76.47	96.98	96.85	97.81		
Other groups of marital status	66.67	97.58	97.39	95.71		
Working condition					9.71	<.001
Working \geq 35 hours	77.78	97.73	97.67	97.69		
Other groups of working condition	68.75	97.05	96.82	95.82		
Race					46.57	<.001
White	79.31	97.07	96.95	97.66		
Black	33.33	97.66	97.36	94.33		
Income					15.38	<.001
An income of <US \$20,000	80.00	94.68	94.54	97.43		
Other groups of income	67.74	97.80	97.63	96.42		

^aAUC: area under the curve.

^bDifference between recall and specificity values.

Figure 5. The receiver operating characteristic (ROC) curves for various groups related to sociodemographic features were used using the neural network model using stochastic gradient descent after matching (with area under the curve [AUC] and 95% CI values). Values were calculated based on the test sample (6583 individuals: 41 developed opioid use disorder [OUD] and 6542 did not develop OUD). (A) ROC curve for sex, (B) ROC curve for marital status, (C) ROC curve for working conditions, (D) ROC curve for race, and (E) ROC curve for income.

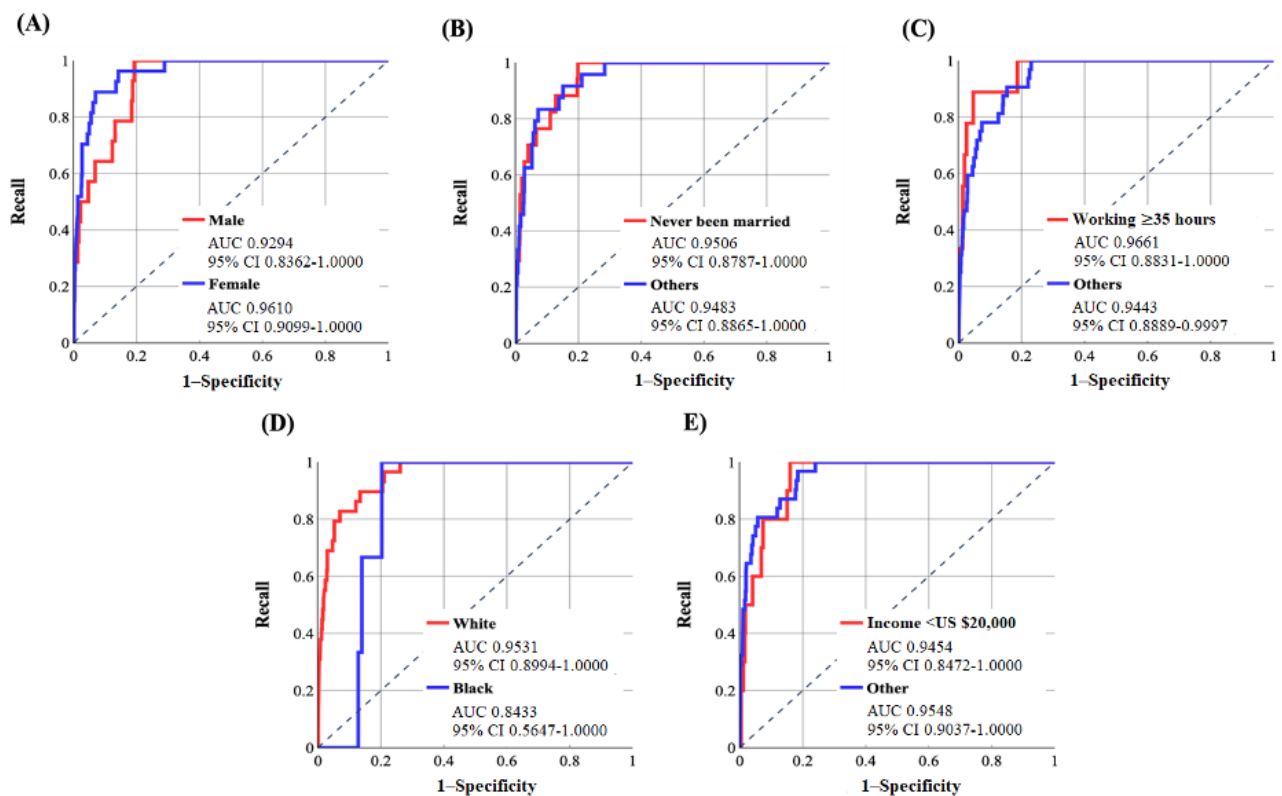


Figure 6. The receiver operating characteristic (ROC) curves for various groups related to sociodemographic features were used using the neural network model using Adam after matching (with area under the curve [AUC] and 95% CI values). Values were calculated based on the test sample (6583 individuals: 41 developed opioid use disorder [OUD] and 6542 did not develop OUD). (A) ROC curve for sex, (B) ROC curve for marital status, (C) ROC curve for working condition, (D) ROC curve for race, and (E) ROC curve for income.

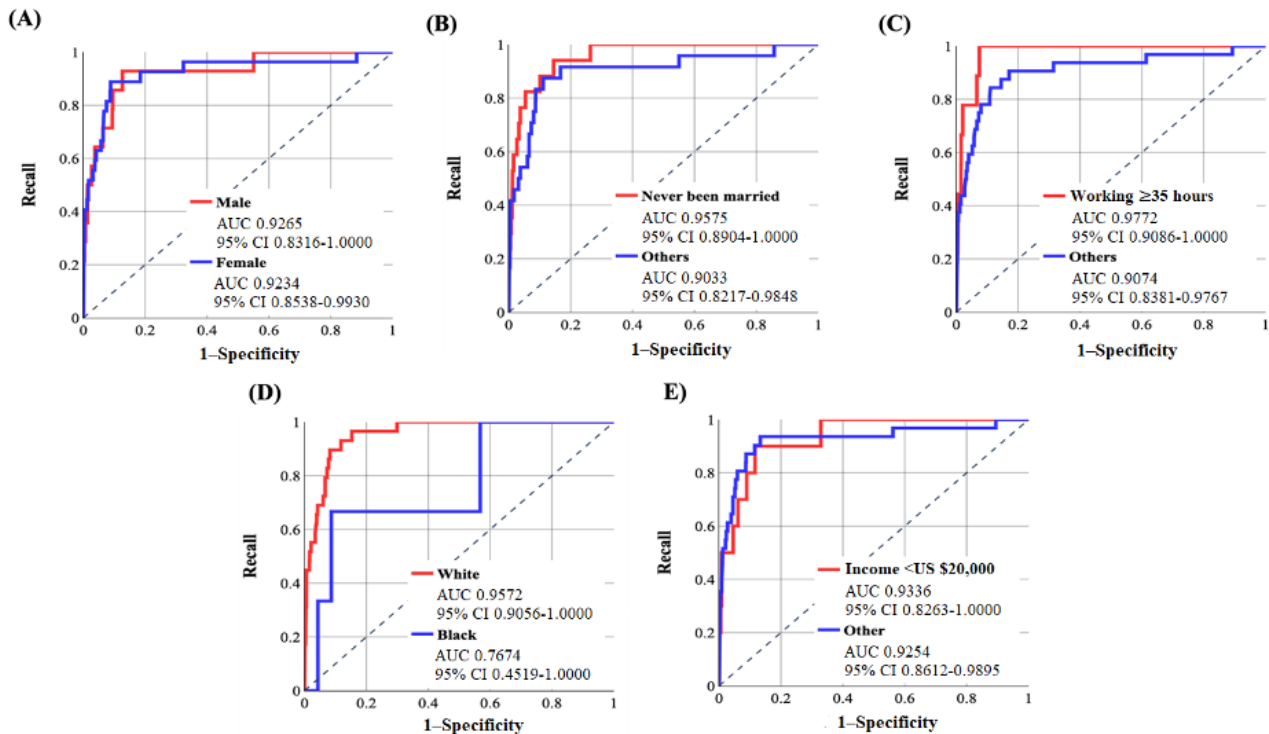


Table 5. The performance metrics of the neural network model using stochastic gradient descent using the default threshold (50%) after matching.

Sociodemographic groups	Recall	Specificity	Accuracy	AUC ^a	Difference ^b	P value
Sex					2.02	<.001
Male	50.00	97.67	97.47	92.94		
Female	51.85	97.84	97.46	96.10		
Marital status					13.02	<.001
Never been married	58.82	97.74	97.54	95.06		
Other groups of marital status	45.83	97.77	97.38	94.83		
Working condition					19.91	<.001
Working ≥ 35 hours	66.67	97.84	97.66	96.61		
Other groups of working condition	46.88	97.72	97.40	94.43		
Race					60.16	<.001
White	58.62	97.46	97.21	95.31		
Black	0.00	99.00	98.51	84.33		
Income					3.42	<.001
An income of <US \$20,000	50.00	96.54	96.33	94.54		
Other groups of income	51.61	98.35	98.02	95.48		

^aAUC: area under the curve.

^bDifference between recall and specificity values.

Table 6. The performance metrics of the neural network model using Adam using the default threshold (50%) after matching.

Sociodemographic groups	Recall	Specificity	Accuracy	AUC ^a	Difference ^b	P value
Sex					2.82	<.001
Male	57.14	96.80	96.63	92.65		
Female	59.26	96.10	95.79	92.34		
Marital status					11.12	<.001
Never been married	64.71	96.74	96.58	95.75		
Other groups of marital status	54.17	96.16	95.84	90.33		
Working condition					25.25	<.001
Working \geq 35 hours	77.78	96.00	95.90	97.72		
Other groups of working condition	53.13	96.60	96.32	90.74		
Race					67.63	<.001
White	65.52	95.90	95.71	95.72		
Black	0.00	98.01	97.52	76.74		
Income					5.40	<.001
An income of <US \$20,000	60.00	94.14	93.99	93.36		
Other groups of income	58.06	97.60	97.32	92.54		

^aAUC: area under the curve.

^bDifference between recall and specificity values.

Bias Detection (Sampling Bias)

Figures 7 and 8 demonstrate the trend of AUC values based on the structure of the training set using the NN-SGD and NN-Adam, respectively.

We observed significant fluctuations in AUC values for all the demographic groups, especially using the NN-SGD, indicating the presence of sampling bias for all sociodemographic features.

The detection of sampling bias for ML classifiers (ie, LR and SVM) is presented in Figures S6-S8 in [Multimedia Appendix 2](#). The LR and linear SVM classifiers demonstrate a significant

sampling bias for all sociodemographic features. In addition, while the SVM-RBF classifier did not show any significant sampling bias for sex, marital status, and working condition, it indicated a notable bias for race and income.

Figures 9 and 10 demonstrate the trend of AUC values based on the structure of the training set using the NN-SGD and NN-Adam after matching, respectively.

We observed significant variations in AUC values, especially using the NN-SGD after matching, highlighting the existence of sampling bias for all sociodemographic features.

Figure 7. The trend of area under the curve (AUC) values for sociodemographic features using the neural network model using stochastic gradient descent: (A) the trend for sex, (B) the trend for marital status, (C) the trend for working conditions, (D) the trend for race, and (E) the trend for income.

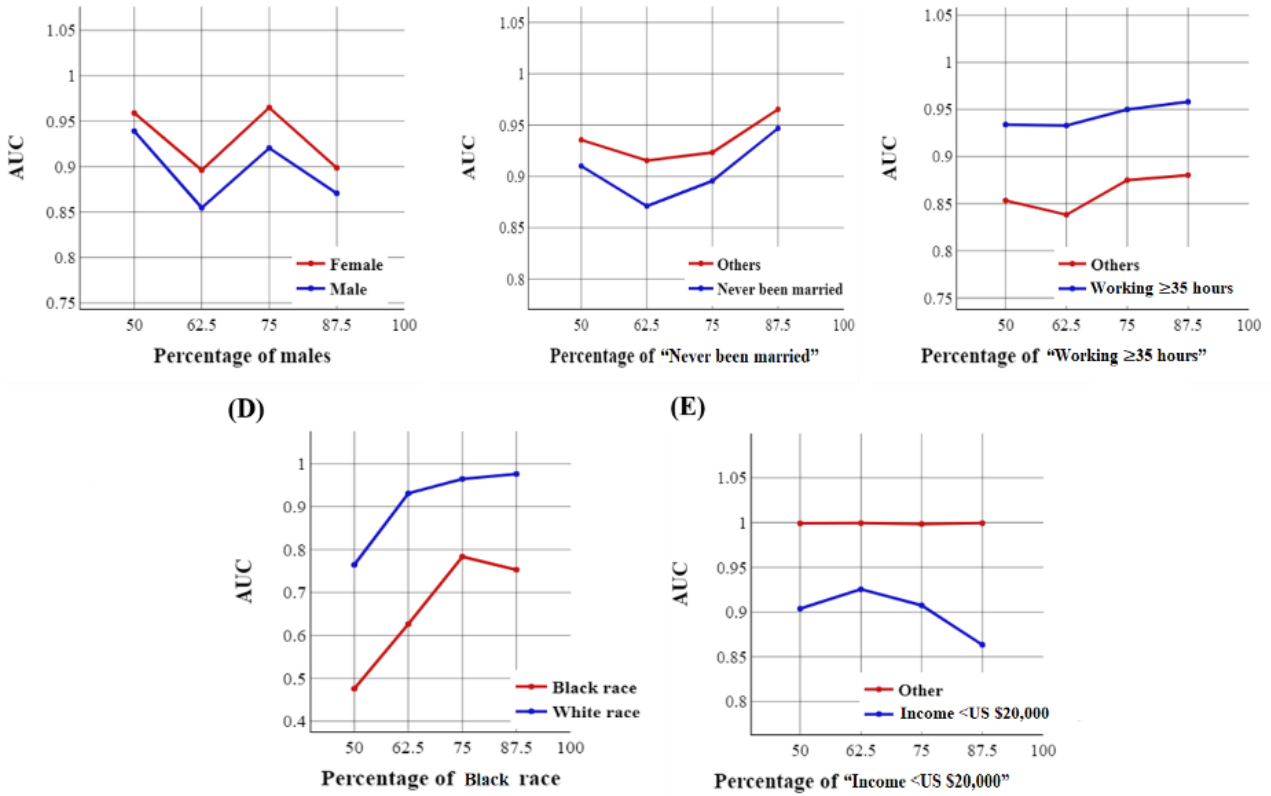


Figure 8. The trend of area under the curve (AUC) values for sociodemographic features using the neural network model using Adam: (A) the trend for sex, (B) the trend for marital status, (C) the trend for working conditions, (D) the trend for race, and (E) the trend for income.

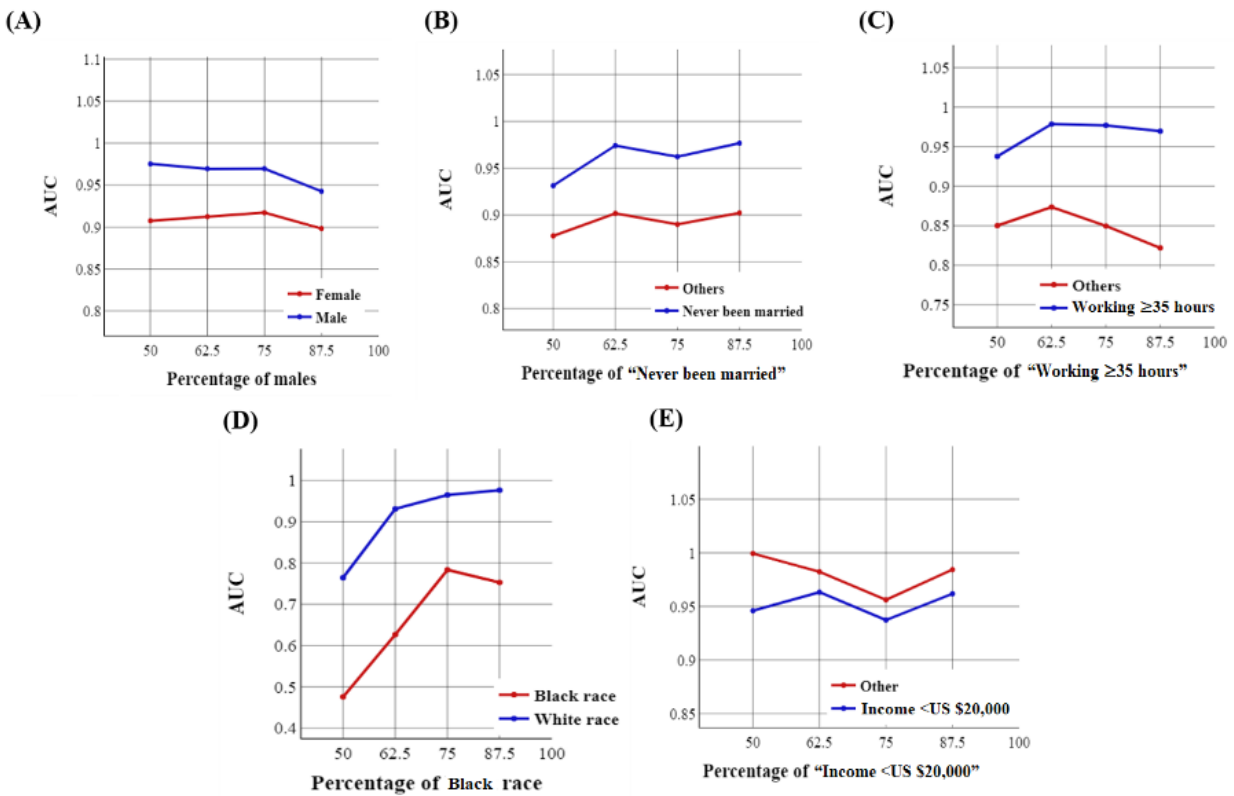


Figure 9. The trend of area under the curve (AUC) values for sociodemographic features using the neural network model using stochastic gradient descent after matching: (A) the trend for sex, (B) the trend for marital status, (C) the trend for working conditions, (D) the trend for race, and (E) the trend for income.

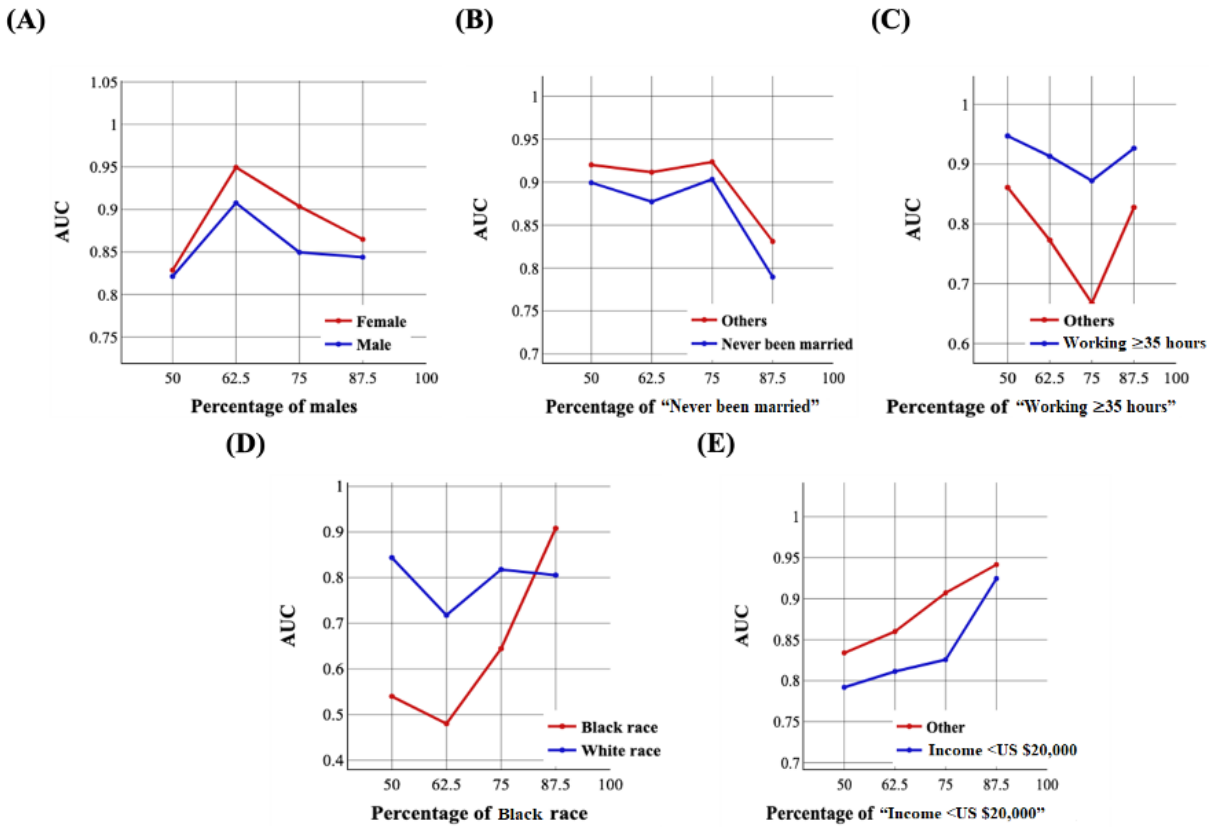
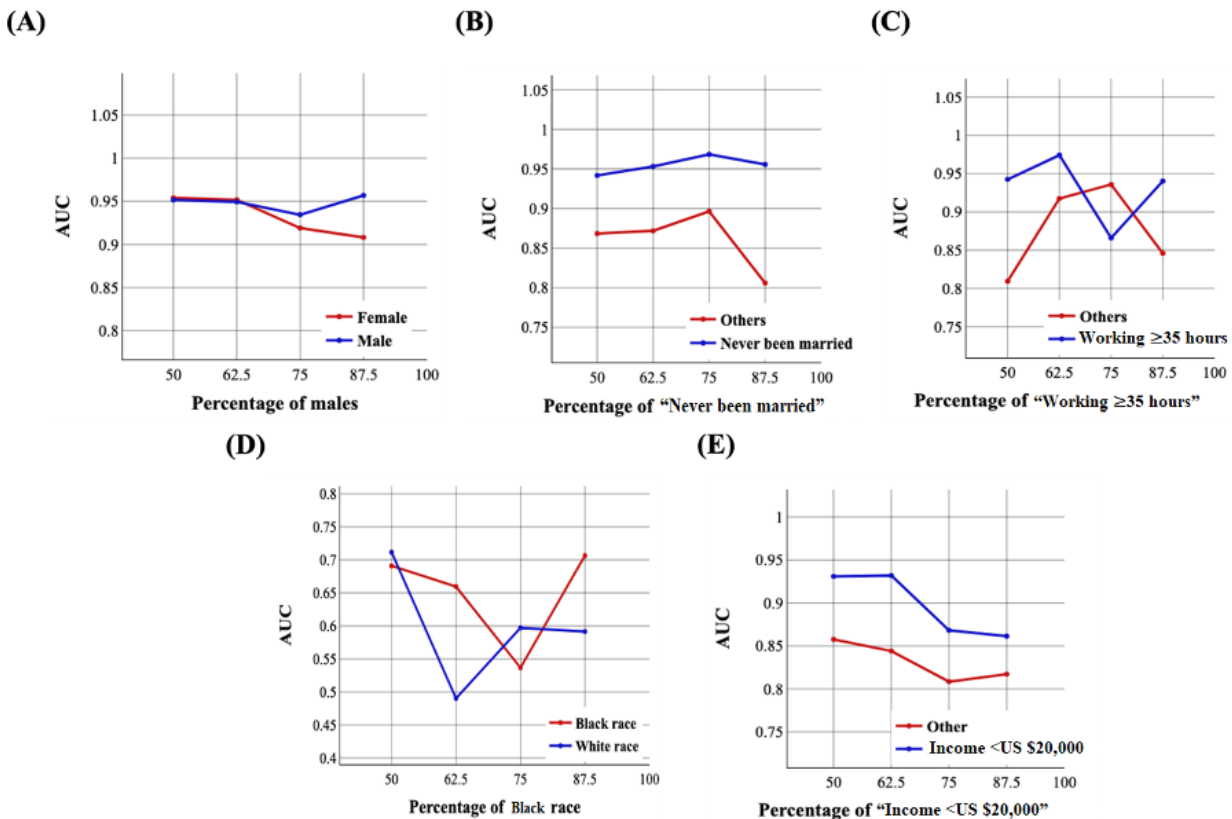


Figure 10. The trend of area under the curve (AUC) values for sociodemographic features using the neural network model using Adam after matching: (A) the trend for sex, (B) the trend for marital status, (C) the trend for working conditions, (D) the trend for race, and (E) the trend for income.



Bias Mitigation

The details of implementing the bias mitigation algorithm, including the optimal threshold and performance metrics, are presented in [Tables 7 and 8](#).

We observed that the recall values for all 5 sociodemographic groups have increased compared to the original NN-SGD using the default threshold of 50%. A similar increase in recall was observed compared to the original NN-Adam, except for *working ≥35 hours* and *income* groups. However, the specificity and accuracy values have decreased for all these groups. Most importantly, the difference between specificity and recall values has decreased for all groups, except the marital status and working condition using the NN-SGD. The reason why the difference has not decreased for the marital status and working condition is that the algorithm enforces the model to have a recall of ≥70% and an accuracy of ≥50%, and thus, it could not find such a threshold after searching all the available options in the range [0,100]. These bias improvements using NN-SGD and NN-Adam for different sociodemographic features are as follows, respectively: sex (21.66% vs 16.96%), marital status (0.00 vs 8.87%), working condition (0.00 vs 8.45%), race (1.48% vs 41.62%), and income (21.04% vs 0.20%).

The improvement in performance of other ML classifiers after implementing our proposed bias mitigation algorithm is presented in [Tables S4-S6 in Multimedia Appendix 2](#). The

results indicate that the algorithm is able to mitigate the bias for all sociodemographic features and improve the recall of the model at the same time. It is notable that using LR and linear SVM classifiers, the recall values did not improve for working condition and income, whereas using the SVM-RBF classifier, the recall values improved for all features. The details regarding the improvements gained by this algorithm are presented in [Table S7 in Multimedia Appendix 2](#).

The details of implementing the bias mitigation algorithm after matching, including the optimal threshold and performance metrics, are presented in [Tables 9 and 10](#).

We observed that the recall values for marital status, working condition, and race have increased compared to the original NN-SGD after matching using the default threshold of 50%. A similar increase in recall was observed compared to the original NN-Adam, except for *working >35 hours* as one of the working condition groups. By contrast, sex and income groups achieved higher specificity and accuracy after matching compared to the original NN-SGD and NN-Adam with a 50% threshold. Notably, the difference between specificity and recall values has decreased for all groups. These bias improvements using NN-SGD and NN-Adam after matching for different sociodemographic features are as follows, respectively: sex (0.14% vs 0.97%), marital status (12.95% vs 10.33%), working condition (14.79% vs 15.33%), race (60.13% vs 41.71%), and income (0.35% vs 2.21%).

Table 7. The details of implementing bias mitigation for the neural network model using stochastic gradient descent.

Sociodemographic conditions	Optimal threshold	Recall	Specificity	Accuracy	Difference ^a
Sex	18.40				1.08 ^b
Male		85.71 ^c	58.99	59.12	
Female		85.19	58.43	58.63	
Marital status	24.10				3.69
Never been married		76.47	87.13	87.07	
Other groups of marital status		75.00	89.35	89.25	
Working condition	24.50				15.22
Working >35 hours		66.67	95.01	94.91	
Other groups of working condition		71.88	85.00	84.90	
Race	18.40				26.60
White		89.66	58.64	58.85	
Black		66.67	62.25	62.27	
Income	17.30				8.89
An income of <US \$20,000		90.00	44.83	45.28	
Other groups of income		87.10	50.82	51.02	

^aDifference between recall and specificity after bias mitigation.

^bItalicized values indicate an improvement compared to the initial values (50% threshold).

Table 8. The details of implementing bias mitigation for the neural network model using Adam.

Sociodemographic conditions	Optimal threshold	Recall	Specificity	Accuracy	Difference ^a
Sex	22.70				4.35 ^b
Male		78.5	95.81	95.73	
Female		81.48	94.37	94.27	
Marital status	0.60				1.53
Never been married		100	65.43	65.64	
Other groups of marital status		100	66.96	67.17	
Working condition	35.40				1.26
Working >35 hours		77.78	96.75	96.68	
Other groups of working condition		78.13	95.84	95.70	
Race	5.90				4.95
White		96.55	90.11	90.15	
Black		100	88.61	88.66	
Income	45.60				15.18
An income of <US \$20,000		80	94.48	94.34	
Other groups of income		67.74	97.40	97.24	

^aDifference between recall and specificity after bias mitigation.

^bItalicized values indicate an improvement compared to the initial values (50% threshold).

Table 9. The details of implementing bias mitigation for the neural network model using stochastic gradient descent after matching.

Sociodemographic conditions	Optimal threshold	Recall	Specificity	Accuracy	Difference ^a
Sex	52.10				1.88 ^b
Male		50.00	97.85	97.65	
Female		51.85	97.88	97.49	
Marital status	5.80				0.07
Never been married		100.00	50.26	50.51	
Other groups of marital status		100.00	50.33	50.69	
Working condition	12.50				5.12
Working ≥35 hours		88.89	89.27	89.27	
Other groups of working condition		87.50	85.54	85.56	
Race	8.10				0.03
White		100.00	73.39	73.56	
Black		100.00	73.42	73.55	
Income	52.60				3.07
An income of <US \$20,000		50.00	96.96	96.74	
Other groups of income		51.61	98.42	98.09	

^aDifference between recall and specificity after bias mitigation.

^bItalicized values indicate an improvement compared with the initial values (50% threshold).

Table 10. The details of implementing bias mitigation for the neural network model using Adam after matching.

Sociodemographic conditions	Optimal threshold	Recall	Specificity	Accuracy	Difference ^a
Sex	66.60				1.85 ^b
Male		50.00	97.88	97.68	
Female		51.85	97.88	97.49	
Marital status	18.60				0.79
Never been married		88.24	88.36	88.36	
Other groups of marital status		87.50	88.31	88.30	
Working condition	33.80				9.92
Working ≥35 hours		77.78	93.33	93.24	
Other groups of working condition		68.75	94.22	94.06	
Race	19.90				25.92
White		89.66	88.27	88.28	
Black		66.67	91.20	91.07	
Income	68.00				3.19
An income of <US \$20,000		50.00	97.00	96.79	
Other groups of income		51.61	98.58	98.25	

^aDifference between recall and specificity after bias mitigation.

^bItalicized values indicate an improvement compared with the initial values (50% threshold).

The Proposed WMV Classifier

As mentioned before, we created a WMV classifier and presented its confusion matrices using the NN-SGD and NN-Adam in Figure 11. The feature weights were calculated based on the difference between recall and specificity values using equation 3 based on the default threshold of 50% (Tables 1 and 2). These weights assigned to NN-SGD and NN-Adam for different features are as follows, respectively: sex (0.23 vs 0.21), marital status (0.03 vs 0.10), working condition (0.15 vs 0.09), race (0.29 vs 0.45), and income (0.30 vs 0.15).

The recall of the WMV classifier is >85% using the NNs trained with both optimizers. In addition, while the specificity and accuracy of this classifier using the NN-SGD are approximately 59%, these values are >90% using the NN-Adam. Compared with the NN-SGD and NN-Adam, the WMV classifier has a significantly higher recall; however, the NNs perform better regarding specificity and accuracy because the WMV classifier uses modified thresholds to mitigate the prediction bias. Overall, this WMV classifier that considers the bias issues for all the sociodemographic features has demonstrated satisfactory performance using the NNs trained with SGD and Adam optimizers and can be used for sufficiently accurate and fairness-aware prediction of OUD in individuals.

The weights assigned to each feature and the confusion matrices of the WMV classifier using the ML classifiers are presented in Table S8 and Figure S9 in Multimedia Appendix 2, respectively. According to the results, the recall values of the

WMV classifier are higher compared to all the original ML classifiers (>92%). Besides, the specificity and accuracy values are sufficiently high for the WMV classifier using all the ML classifiers (>75%).

Figure 12 shows the confusion matrices of the WMV classifier using the NN-SGD and NN-Adam after matching. The feature weights were calculated based on the difference between recall and specificity values using equation 3 based on the default threshold of 50% (Tables 3 and 4). These weights assigned to NN-SGD and NN-Adam for different features are as follows, respectively: sex (0.02 vs 0.03), marital status (0.13 vs 0.10), working condition (0.20 vs 0.23), race (0.61 vs 0.60), and income (0.03 vs 0.05).

The recall of the WMV classifier is >85% using the NNs trained with both optimizers. In addition, while the specificity and accuracy of this classifier using the NN-SGD are approximately 73%, these values are >89% using the NN-Adam. Compared to the NN-SGD and NN-Adam, the WMV classifier has a significantly higher recall; however, the NNs have higher specificity and accuracy because the WMV classifier uses thresholds for bias mitigation. Overall, this WMV classifier can be used as a fairness-aware predictor of OUD in real-world applications, guiding clinicians in fair and accurate decision-making.

Table 11 demonstrates the performance of different models, including NN-SGD, NN-Adam, and WMV classifiers before and after 1-N matching.

Figure 11. The confusion matrix of the weighted majority voting classifier using neural networks (NNs): (A) NN model using stochastic gradient descent and (B) NN model using Adam.

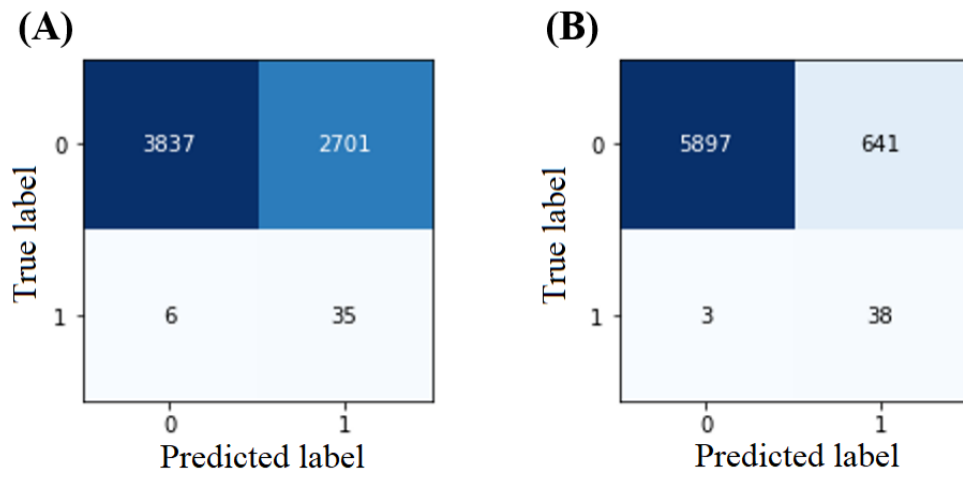


Figure 12. The confusion matrix of the weighted majority voting classifier using neural networks (NNs) after matching: (A) NN model using stochastic gradient descent and (B) NN model using Adam.

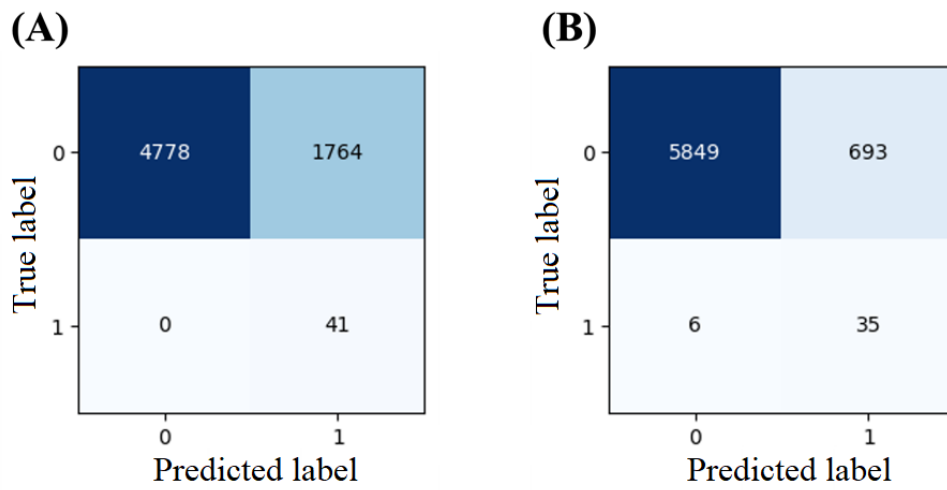


Table 11. The performance of different classifiers before and after matching.

Status	Recall	Specificity	Accuracy	Precision
NN-SGD^a				
Before matching	21.95	99.46	98.98	20.45
After matching	51.22	97.75	97.46	12.50
NN-Adam^b				
Before matching	70.73	97.32	97.16	14.22
After matching	58.54	96.45	96.22	9.38
WMV^c (NN-SGD)				
Before matching	85.37	41.31	58.85	1.28
After matching	100.00	73.04	73.20	2.27
WMV (NN-Adam)				
Before matching	92.68	90.20	90.21	5.60
After matching	85.37	89.41	89.38	4.81

^aNN-SGD: neural network model using stochastic gradient descent.

^bNN-Adam: neural network model using Adam.

^cWMV: weighted majority voting.

Discussion

Principal Findings

According to the results, the proposed bias mitigation algorithm performs well in reducing the bias and producing fairer results for individuals in different sociodemographic groups. However, there is always a trade-off between the bias and the accuracy and specificity of the model. Although the recall values have improved for all sociodemographic groups and the bias has been remarkably mitigated, the accuracy and specificity values have dropped for all these features. Notably, although we could mitigate the bias to a larger extent by solely changing the threshold to minimize the difference, we applied threshold values for accuracy (50%) and recall (70%) so that the overall performance of the model would remain satisfactory. This depends on the user preferences and the importance of bias compared to the model performance in a real-world setting.

The achievements of the proposed bias mitigation algorithm and WMV classifier represent a significant advancement in the field of ML for health care, particularly in addressing fairness and equity in predicting OUD. In the context of health applications, where demographic disparities can lead to unequal treatment outcomes, the ability of this algorithm to substantially reduce bias while enhancing recall is crucial. By trying to equalize recall and specificity across all sociodemographic groups, the algorithm ensures that individuals at risk are equally identified across the groups, which is vital for early and fair intervention and treatment. Compared to existing methods [39,63,64], this approach offers a more rigorous solution considering the performance threshold for both accuracy and recall at the same time. Thus, the classifier maintains an overall satisfactory performance, which is essential in real-world clinical settings where both fairness and accuracy are critical for patient outcomes and sacrificing the performance for achieving fairness

is not desirable. These improvements provide a clearer understanding of the impact of the work in real-world health care applications. In addition, unlike many methods that might only mitigate bias for a single feature at a time [39,63,64], our proposed approach mitigates bias for all sociodemographic features, and then, all the results are incorporated into a WMV classifier, making it more viable for deployment in diverse health care environments. Overall, the proposed algorithm and classifier represent a meaningful step forward in creating fairer and more effective ML models for predicting outcomes, such as OUD, thereby potentially improving health equity and treatment efficacy in clinical practice.

The application of the proposed bias mitigation algorithm could be extended far beyond the OUD prediction presented in this study. For example, in the realm of racial bias, the algorithm can be applied to predictive models for cardiovascular disease, ensuring that both Black and White patients receive both equal and accurate risk assessments, thereby improving early detection and treatment for Black patients who might otherwise be overlooked [64]. Similarly, Black women experience a 3 times higher likelihood of mortality from pregnancy-related causes compared to their White counterparts [65], where a biased model could underdiagnose or misdiagnose African Americans, leading to inadequate treatment and poorer health outcomes. In addressing sex bias in heart disease prediction, the algorithm can adjust thresholds to enhance both equality and accuracy for women, ensuring their symptoms are not dismissed and they receive timely care [64]. Moreover, in tackling socioeconomic bias, the algorithm can be used in models predicting the risk of chronic disease, ensuring that individuals from lower-income backgrounds are equally and accurately assessed, leading to equitable health care interventions [66]. In the case of diabetes, where African American populations have higher rates of diabetes compared to non-Hispanic White individuals [67], a biased predictive model might fail to identify at-risk individuals

in these minority groups, resulting in delayed diagnosis and treatment. Furthermore, women are more likely to be diagnosed with depression because of societal and sex norms [68], necessitating bias mitigation in predictive models. By addressing these biases while preserving a high performance, the proposed algorithm promotes fairness and improves the reliability of health outcomes predictions across diverse patient populations, making it a valuable tool for enhancing health equity and preventing complications such as cardiovascular issues and eventually death.

After implementing the bias mitigation algorithm, we proposed the WMV classifier to classify the inputs based on the proposed thresholds for sex, marital status, working condition, race, and income. Although this classifier works based on the classification performed for each feature, one may prefer to use the threshold for sex if it is more important than other sociodemographic features in a particular study. This could be the case for other features as well, depending on the user preferences and the type and nature of the study being performed.

In this study, we used SGD and Adam to train the NN models. The reason why we used these 2 optimizers is that we could obtain different bias mitigation results. For example, although the NN-SGD could perform better in mitigating the sex bias compared to NN-Adam (21.66% vs 19.96%), the NN-Adam better mitigated the race bias (41.62% vs 1.48%). Accordingly, one could prefer to use a specific hyperparameter based on the nature of the study and user preferences. For instance, if sex is a more important feature than race in a certain case, Adam optimizer would be a better choice compared to SGD. Overall, we chose the NN-Adam as the final best classifier as it could mitigate the bias for all 5 sociodemographic features and have a higher predictive performance after developing the WMV classifier.

In this study, we observed that the precision is lower than the recall in the WMV classifier. While precision is an important metric that indicates the proportion of true positive predictions among all positive predictions, recall holds higher importance and utility in clinical decision-making [69]. Recall measures the ability of the model to correctly identify all relevant cases, in this instance, true positives among those who actually have OUD. Missing a true positive (ie, a false negative) can have severe consequences, potentially delaying critical treatments or interventions. Therefore, a higher recall ensures that most patients with the condition are identified, even in the presence of high false positives. This trade-off is critical in clinical practice, where the cost of misclassifying a patient at high risk of OUD outweighs the cost of additional testing or follow-up for misclassifying an actual non-OUD patient. Hence, despite the lower precision, the higher recall of our model provides greater overall utility in ensuring patient safety and effective clinical outcomes.

We analyzed the effectiveness of the proposed algorithm using several ML models, including LR, linear SVM, and SVM-RBF. Other ML models exist, such as random forests and decision trees, which could potentially classify OUD with high predictive performance. However, these models do not assign probability

values to the output classes, and the proposed algorithm cannot be used to mitigate their potential bias.

Limitations

This study used the 2020 NSDUH data, with most cases belonging to the non-OUD class and <1% to the OUD class. While it is important to acknowledge this data limitation and contextualize it within the body of research, we used 2 techniques, including class weighting and 1-N matching, to address the class imbalance problem. The experiments following the 1-N matching demonstrated the existence of bias, which was mitigated using the proposed bias mitigation algorithm. Moreover, it is notable that several previous studies have successfully applied ML to predict OUD [5,16,18,19]. Despite being highly imbalanced and including much less positive OUD cases than negative ones, many studies have demonstrated the remarkable potential of ML models for OUD prediction [19]. For example, Hasan et al [18] and Lo-Ciganic et al [16] used credible, real-world claims data to predict OUD with high performance despite their notably high imbalance. Similarly, Han et al [5] used NSDUH data to predict OUD among the US population. These studies demonstrate that, although the data are significantly imbalanced, ML can be effectively used to predict OUD, providing valuable insights and aiding in early intervention strategies.

Although the proposed algorithm works well in removing the bias, some limitations exist in this study. The algorithm can noticeably reduce the bias for a single variable (such as sex) and propose an optimal threshold. However, it cannot suggest a single threshold that best mitigates the bias for a group of variables. Moreover, although we demonstrated the existence of bias for demographic features with multiple groups, our algorithm can consider only 2 different groups at the same time. Furthermore, although we included race as a sociodemographic feature, the number of individuals belonging to the Black race who had developed OUD was very low compared to the White race (3 vs 29), which could degrade the generalization of the classifiers. Therefore, including more individuals from the Black race could improve the reliability of the classifiers in real-world applications.

The proposed bias mitigation algorithm, although effective in reducing bias for OUD prediction, can introduce new forms of bias or overlook specific subpopulations. For instance, within racial categories, specific ethnic subgroups, such as Native Americans or recent immigrants, could be overlooked. These groups might have unique cultural or socioeconomic factors affecting their risk of OUD, leading to biased outcomes if these variations are not captured [70]. Similarly, young adults and older adults might experience OUD differently because of distinct life stages and associated risk factors. Young adults might be more susceptible to peer pressure and experimental substance use, while older adults may have chronic pain issues, leading to prolonged opioid prescriptions [71]. If the model does not adequately capture these age-specific differences, predictions could be less accurate for these groups. In addition, certain groups considered vulnerable such as individuals who have been incarcerated or those experiencing homelessness might not be adequately represented in the survey data. These

populations often have higher rates of substance use disorders and face different risk factors compared to the general population. Their exclusion or underrepresentation can result in a model that does not generalize well to these groups, leading to biased predictions. Despite these potential drawbacks, the model offers significant advantages. Systematic adjustment of classification thresholds for existing sociodemographic features ensures balanced predictive performance across different demographic groups, reducing discrimination and improving fairness.

The development and implementation of the WMV classifier enhances the applicability of the proposed bias mitigation algorithm, allowing for tailored threshold adjustments based on the importance of specific sociodemographic features in different studies. This flexibility ensures that the algorithm can be adapted to various contexts, addressing specific fairness concerns as needed. While the WMV classifier performed well in the accurate prediction of OUD, its reliance on proportional importance might not fully capture real-world complexities, such as the interplay between various sociodemographic and health factors. For example, the importance of income might be overestimated, ignoring how low socioeconomic status intersects with other factors such as access to health care and social support networks, thus affecting the model's accuracy for people from different socioeconomic backgrounds. Educational attainment, geographic location, employment status, occupation types, and housing stability also influence OUD risk and may not be fully accounted for, potentially skewing results and introducing new biases. Continuous evaluation and refinement are necessary to ensure that the model addresses these complexities, minimizing new biases and ensuring equitable outcomes across all populations.

Conclusions

The OUD is the result of irregular opioid use, which is a significant cause of deaths worldwide. The ML models have great potential in OUD prediction; however, these models are prone to bias because of the existence of sociodemographic features. In this study, we proposed a bias mitigation algorithm based on EO. This algorithm works based on the threshold moving to achieve an optimal threshold, minimizing the

difference between the specificity and recall values for sociodemographic groups. In addition, this algorithm considers the threshold for the overall recall and accuracy to ensure that the model performs well in OUD prediction. Finally, we proposed a WMV classifier that makes predictions based on the optimal thresholds for all sociodemographic features. The results suggest that the proposed algorithm achieves 21.66%, 1.48%, and 21.04% bias improvement for sex, race, and income using the NN-SGD. The algorithm using the NN-Adam shows an improvement of 16.96%, 8.87%, 8.45%, 41.62%, and 0.20% for sex, marital status, working condition, race, and income, respectively. This algorithm was also able to increase the recall of these classifiers at the same time. In addition, the WMV classifier achieved recall values of 85.37% and 92.68%, specificity values of 58.69% and 90.20%, and accuracy values of 58.85% and 90.21% using NN-SGD and NN-Adam, respectively. This WMV classifier has the potential to be used as a fairness-aware OUD predictor in a real-world setting. The results of the proposed bias mitigation algorithm and WMV classifier for 3 ML classifiers, including LR, linear SVM, and SVM-RBF, also prove the effectiveness of these methods in bias mitigation and fairness-aware prediction of OUD.

Although this study has achieved its research goals, the recommendations for future research work are as follows. First, the bias mitigation algorithm can be extended by developing a method that considers groups of sociodemographic variables and suggests an optimal global threshold. Second, the algorithm can be extended by developing an approach for mitigating the bias and selecting a threshold value for multigroup sociodemographic features instead of focusing on 2 groups simultaneously. Third, the performance of the bias mitigation algorithm may improve by training the NNs with different hyperparameters, such as the learning rate and optimizer. Fourth, more balanced data containing a higher proportion of samples belonging to the minority class and other sociodemographic features can be used to develop fairness-aware predictive models for real-world applications. Fifth, the proposed methods can be used in other medical applications, including but not limited to disease detection, disease classification, and treatment response prediction.

Conflicts of Interest

None declared.

Multimedia Appendix 1

The details of variables used in the study.

[\[DOCX File, 31 KB - ai_v3i1e55820_app1.docx\]](#)

Multimedia Appendix 2

The performance of the proposed bias mitigation algorithm using machine learning classifiers.

[\[DOCX File, 1067 KB - ai_v3i1e55820_app2.docx\]](#)

References

1. Hasan MM, Faiz TI, Modestino AS, Young GJ, Noor-E-Alam M. Optimizing return and secure disposal of prescription opioids to reduce the diversion to secondary users and black market. *Socio Econ Plan Sci* 2023 Apr;86:101457 [FREE Full text] [doi: [10.1016/j.seps.2022.101457](https://doi.org/10.1016/j.seps.2022.101457)]
2. U.S. overdose deaths in 2021 increased half as much as in 2020 – but are still up 15%. Centers for Disease Control and Prevention. URL: https://www.cdc.gov/nchs/pressroom/nchs_press_releases/2022/202205.htm [accessed 2024-04-29]
3. Understanding the opioid overdose epidemic. Centers for Disease Control and Prevention. URL: <https://www.cdc.gov/opioids/basics/epidemic.html> [accessed 2024-04-29]
4. Xia Z, Stewart K. A counterfactual analysis of opioid-involved deaths during the COVID-19 pandemic using a spatiotemporal random forest modeling approach. *Health Place* 2023 Mar;80:102986 [FREE Full text] [doi: [10.1016/j.healthplace.2023.102986](https://doi.org/10.1016/j.healthplace.2023.102986)] [Medline: [36774811](https://pubmed.ncbi.nlm.nih.gov/36774811/)]
5. Han DH, Lee S, Seo DC. Using machine learning to predict opioid misuse among U.S. adolescents. *Prev Med* 2020 Jan;130:105886. [doi: [10.1016/j.ypmed.2019.105886](https://doi.org/10.1016/j.ypmed.2019.105886)] [Medline: [31705938](https://pubmed.ncbi.nlm.nih.gov/31705938/)]
6. Ciesielski T, Iyengar R, Bothra A, Tomala D, Cislo G, Gage BF. A tool to assess risk of de novo opioid abuse or dependence. *Am J Med* 2016 Jul;129(7):699-705.e4 [FREE Full text] [doi: [10.1016/j.amjmed.2016.02.014](https://doi.org/10.1016/j.amjmed.2016.02.014)] [Medline: [26968469](https://pubmed.ncbi.nlm.nih.gov/26968469/)]
7. Dufour R, Mardekian J, Pasquale M, Schaaf D, Andrews GA, Patel NC. Understanding predictors of opioid abuse: predictive model development and validation. *Am J Pharm Benefits* 2014;6:208-216 [FREE Full text]
8. Edlund MJ, Martin BC, Fan MY, Devries A, Braden JB, Sullivan MD. Risks for opioid abuse and dependence among recipients of chronic opioid therapy: results from the TROUP study. *Drug Alcohol Depend* 2010 Nov 01;112(1-2):90-98 [FREE Full text] [doi: [10.1016/j.drugalcdep.2010.05.017](https://doi.org/10.1016/j.drugalcdep.2010.05.017)] [Medline: [20634006](https://pubmed.ncbi.nlm.nih.gov/20634006/)]
9. Hylan TR, Von Korff M, Saunders K, Masters E, Palmer RE, Carrell D, et al. Automated prediction of risk for problem opioid use in a primary care setting. *J Pain* 2015 Apr;16(4):380-387 [FREE Full text] [doi: [10.1016/j.jpain.2015.01.011](https://doi.org/10.1016/j.jpain.2015.01.011)] [Medline: [25640294](https://pubmed.ncbi.nlm.nih.gov/25640294/)]
10. Ives TJ, Chelminski PR, Hammett-Stabler CA, Malone RM, Perhac JS, Potisek NM, et al. Predictors of opioid misuse in patients with chronic pain: a prospective cohort study. *BMC Health Serv Res* 2006 Apr 04;6:46 [FREE Full text] [doi: [10.1186/1472-6963-6-46](https://doi.org/10.1186/1472-6963-6-46)] [Medline: [16595013](https://pubmed.ncbi.nlm.nih.gov/16595013/)]
11. Rice JB, White AG, Birnbaum HG, Schiller M, Brown DA, Roland CL. A model to identify patients at risk for prescription opioid abuse, dependence, and misuse. *Pain Med* 2012 Sep 01;13(9):1162-1173. [doi: [10.1111/j.1526-4637.2012.01450.x](https://doi.org/10.1111/j.1526-4637.2012.01450.x)] [Medline: [22845054](https://pubmed.ncbi.nlm.nih.gov/22845054/)]
12. White AG, Birnbaum HG, Schiller M, Tang J, Katz NP. Analytic models to identify patients at risk for prescription opioid abuse. *Am J Manag Care* 2009 Dec;15(12):897-906 [FREE Full text] [Medline: [20001171](https://pubmed.ncbi.nlm.nih.gov/20001171/)]
13. Turk DC, Swanson KS, Gatchel RJ. Predicting opioid misuse by chronic pain patients: a systematic review and literature synthesis. *Clin J Pain* 2008;24(6):497-508. [doi: [10.1097/AJP.0b013e31816b1070](https://doi.org/10.1097/AJP.0b013e31816b1070)] [Medline: [18574359](https://pubmed.ncbi.nlm.nih.gov/18574359/)]
14. Thornton JD, Dwibedi N, Scott V, Ponte CD, Ziedonis D, Sambamoorthi N, et al. Predictors of transitioning to incident chronic opioid therapy among working-age adults in the United States. *Am Health Drug Benefits* 2018 Feb;11(1):12-21 [FREE Full text] [Medline: [29692877](https://pubmed.ncbi.nlm.nih.gov/29692877/)]
15. Skala K, Reichl L, Ilias W, Likar R, Groggl-Aringer G, Wallner C, et al. Can we predict addiction to opioid analgesics? A possible tool to estimate the risk of opioid addiction in patients with pain. *Pain Physician* 2013;16(6):593-601 [FREE Full text] [Medline: [24284844](https://pubmed.ncbi.nlm.nih.gov/24284844/)]
16. Lo-Ciganic WH, Huang JL, Zhang HH, Weiss JC, Wu Y, Kwok CK, et al. Evaluation of machine-learning algorithms for predicting opioid overdose risk among Medicare beneficiaries with opioid prescriptions. *JAMA Netw Open* 2019 Mar 01;2(3):e190968 [FREE Full text] [doi: [10.1001/jamanetworkopen.2019.0968](https://doi.org/10.1001/jamanetworkopen.2019.0968)] [Medline: [30901048](https://pubmed.ncbi.nlm.nih.gov/30901048/)]
17. Chiew CJ, Liu N, Wong TH, Sim YE, Abdullah HR. Utilizing machine learning methods for preoperative prediction of postsurgical mortality and intensive care unit admission. *Ann Surg* 2020 Dec;272(6):1133-1139 [FREE Full text] [doi: [10.1097/SLA.0000000000003297](https://doi.org/10.1097/SLA.0000000000003297)] [Medline: [30973386](https://pubmed.ncbi.nlm.nih.gov/30973386/)]
18. Hasan MM, Young GJ, Patel MR, Modestino AS, Sanchez LD, Noor-E-Alam M. A machine learning framework to predict the risk of opioid use disorder. *Mach Learn Appl* 2021 Dec;6:100144 [FREE Full text] [doi: [10.1016/j.mlwa.2021.100144](https://doi.org/10.1016/j.mlwa.2021.100144)]
19. Garbin C, Marques N, Marques O. Machine learning for predicting opioid use disorder from healthcare data: a systematic review. *Comput Methods Programs Biomed* 2023 Jun;236:107573. [doi: [10.1016/j.cmpb.2023.107573](https://doi.org/10.1016/j.cmpb.2023.107573)] [Medline: [37148670](https://pubmed.ncbi.nlm.nih.gov/37148670/)]
20. Tseregounis IE, Henry SG. Assessing opioid overdose risk: a review of clinical prediction models utilizing patient-level data. *Transl Res* 2021 Aug;234:74-87 [FREE Full text] [doi: [10.1016/j.trsl.2021.03.012](https://doi.org/10.1016/j.trsl.2021.03.012)] [Medline: [33762186](https://pubmed.ncbi.nlm.nih.gov/33762186/)]
21. Ellis RJ, Wang Z, Genes N, Ma'ayan A. Predicting opioid dependence from electronic health records with machine learning. *BioData Min* 2019 Jan 29;12(1):3 [FREE Full text] [doi: [10.1186/s13040-019-0193-0](https://doi.org/10.1186/s13040-019-0193-0)] [Medline: [30728857](https://pubmed.ncbi.nlm.nih.gov/30728857/)]
22. Wadekar AS. Understanding opioid use disorder (OUD) using tree-based classifiers. *Drug Alcohol Depend* 2020 Mar 01;208:107839. [doi: [10.1016/j.drugalcdep.2020.107839](https://doi.org/10.1016/j.drugalcdep.2020.107839)] [Medline: [31962227](https://pubmed.ncbi.nlm.nih.gov/31962227/)]
23. Lo-Ciganic WH, Donohue JM, Yang Q, Huang JL, Chang C, Weiss JC, et al. Developing and validating a machine-learning algorithm to predict opioid overdose in Medicaid beneficiaries in two US states: a prognostic modelling study. *Lancet Digit Health* 2022 Jun;4(6):e455-e465 [FREE Full text] [doi: [10.1016/S2589-7500\(22\)00062-0](https://doi.org/10.1016/S2589-7500(22)00062-0)] [Medline: [35623798](https://pubmed.ncbi.nlm.nih.gov/35623798/)]

24. Dong X, Deng J, Rashidian S, Abell-Hart K, Hou W, Rosenthal RN, et al. Identifying risk of opioid use disorder for patients taking opioid medications with deep learning. *J Am Med Inform Assoc* 2021 Jul 30;28(8):1683-1693 [FREE Full text] [doi: [10.1093/jamia/ocab043](https://doi.org/10.1093/jamia/ocab043)] [Medline: [33930132](https://pubmed.ncbi.nlm.nih.gov/33930132/)]
25. Warren D, Marashi A, Siddiqui A, Eijaz AA, Pradhan P, Lim D, et al. Using machine learning to study the effect of medication adherence in Opioid use disorder. *PLoS One* 2022 Dec 15;17(12):e0278988 [FREE Full text] [doi: [10.1371/journal.pone.0278988](https://doi.org/10.1371/journal.pone.0278988)] [Medline: [36520864](https://pubmed.ncbi.nlm.nih.gov/36520864/)]
26. Annis IE, Jordan R, Thomas KC. Quickly identifying people at risk of opioid use disorder in emergency departments: trade-offs between a machine learning approach and a simple EHR flag strategy. *BMJ Open* 2022 Sep 14;12(9):e059414 [FREE Full text] [doi: [10.1136/bmjopen-2021-059414](https://doi.org/10.1136/bmjopen-2021-059414)] [Medline: [36104124](https://pubmed.ncbi.nlm.nih.gov/36104124/)]
27. Afshar M, Sharma B, Bhalla S, Thompson HM, Dligach D, Boley RA, et al. External validation of an opioid misuse machine learning classifier in hospitalized adult patients. *Addict Sci Clin Pract* 2021 Mar 17;16(1):19 [FREE Full text] [doi: [10.1186/s13722-021-00229-7](https://doi.org/10.1186/s13722-021-00229-7)] [Medline: [33731210](https://pubmed.ncbi.nlm.nih.gov/33731210/)]
28. Dong X, Wong R, Lyu W, Abell-Hart K, Deng J, Liu Y, et al. An integrated LSTM-HeteroRGNN model for interpretable opioid overdose risk prediction. *Artif Intell Med* 2023 Jan;135:102439 [FREE Full text] [doi: [10.1016/j.artmed.2022.102439](https://doi.org/10.1016/j.artmed.2022.102439)] [Medline: [36628797](https://pubmed.ncbi.nlm.nih.gov/36628797/)]
29. Dong X, Deng J, Hou W, Rashidian S, Rosenthal RN, Saltz M, et al. Predicting opioid overdose risk of patients with opioid prescriptions using electronic health records based on temporal deep learning. *J Biomed Inform* 2021 Apr;116:103725 [FREE Full text] [doi: [10.1016/j.jbi.2021.103725](https://doi.org/10.1016/j.jbi.2021.103725)] [Medline: [33711546](https://pubmed.ncbi.nlm.nih.gov/33711546/)]
30. Sanger N, Bhatt M, Shams I, Shahid H, Luo C, Tam SL, et al. Association between socio-demographic and health functioning variables among patients with opioid use disorder introduced by prescription: a prospective cohort study. *Pain Physician* 2018 Nov;21(6):E623-E632 [FREE Full text] [Medline: [30508993](https://pubmed.ncbi.nlm.nih.gov/30508993/)]
31. Parlier-Ahmad AB, Martin CE, Radic M, Svikis DS. An exploratory study of sex and gender differences in demographic, psychosocial, clinical, and substance use treatment characteristics of patients in outpatient opioid use disorder treatment with buprenorphine. *Transl Issues Psychol Sci* 2021 Jun;7(2):141-153 [FREE Full text] [doi: [10.1037/tps0000250](https://doi.org/10.1037/tps0000250)] [Medline: [34541257](https://pubmed.ncbi.nlm.nih.gov/34541257/)]
32. Altekruse SF, Cosgrove CM, Altekruse WC, Jenkins RA, Blanco C. Socioeconomic risk factors for fatal opioid overdoses in the United States: findings from the mortality disparities in American Communities Study (MDAC). *PLoS One* 2020;15(1):e0227966 [FREE Full text] [doi: [10.1371/journal.pone.0227966](https://doi.org/10.1371/journal.pone.0227966)] [Medline: [31951640](https://pubmed.ncbi.nlm.nih.gov/31951640/)]
33. Lee CW, Lo YT, Devi S, Seo Y, Simon A, Zborovancik K, et al. Gender differences in preoperative opioid use in spine surgery patients: a systematic review and meta-analysis. *Pain Med* 2020 Dec 25;21(12):3292-3300. [doi: [10.1093/pm/pnaa266](https://doi.org/10.1093/pm/pnaa266)] [Medline: [32989460](https://pubmed.ncbi.nlm.nih.gov/32989460/)]
34. Back SE, Payne RL, Wahlquist AH, Carter RE, Stroud Z, Haynes L, et al. Comparative profiles of men and women with opioid dependence: results from a national multisite effectiveness trial. *Am J Drug Alcohol Abuse* 2011 Sep 22;37(5):313-323 [FREE Full text] [doi: [10.3109/00952990.2011.596982](https://doi.org/10.3109/00952990.2011.596982)] [Medline: [21854273](https://pubmed.ncbi.nlm.nih.gov/21854273/)]
35. Olsen Y, Daumit GL, Ford DE. Opioid prescriptions by U.S. primary care physicians from 1992 to 2001. *J Pain* 2006 Apr;7(4):225-235 [FREE Full text] [doi: [10.1016/j.jpain.2005.11.006](https://doi.org/10.1016/j.jpain.2005.11.006)] [Medline: [16618466](https://pubmed.ncbi.nlm.nih.gov/16618466/)]
36. Pletcher MJ, Kertesz SG, Kohn MA, Gonzales R. Trends in opioid prescribing by race/ethnicity for patients seeking care in US emergency departments. *JAMA* 2008 Jan 02;299(1):70-78. [doi: [10.1001/jama.2007.64](https://doi.org/10.1001/jama.2007.64)] [Medline: [18167408](https://pubmed.ncbi.nlm.nih.gov/18167408/)]
37. Anderson KO, Green CR, Payne R. Racial and ethnic disparities in pain: causes and consequences of unequal care. *J Pain* 2009 Dec;10(12):1187-1204 [FREE Full text] [doi: [10.1016/j.jpain.2009.10.002](https://doi.org/10.1016/j.jpain.2009.10.002)] [Medline: [19944378](https://pubmed.ncbi.nlm.nih.gov/19944378/)]
38. Fahse T, Huber V, van Giffen B. Managing bias in machine learning projects. In: Ahlemann F, Schütte R, Stieglitz S, editors. *Innovation Through Information Systems: Volume II: A Collection of Latest Research on Technology Issues*. Cham, Switzerland: Springer; 2021:94-109.
39. Fletcher RR, Nakeshimana A, Olubeko O. Addressing fairness, bias, and appropriate use of artificial intelligence and machine learning in global health. *Front Artif Intell* 2020;3:561802 [FREE Full text] [doi: [10.3389/frai.2020.561802](https://doi.org/10.3389/frai.2020.561802)] [Medline: [33981989](https://pubmed.ncbi.nlm.nih.gov/33981989/)]
40. Gianfrancesco MA, Tamang S, Yazdany J, Schmajuk G. Potential biases in machine learning algorithms using electronic health record data. *JAMA Intern Med* 2018 Nov 01;178(11):1544-1547 [FREE Full text] [doi: [10.1001/jamainternmed.2018.3763](https://doi.org/10.1001/jamainternmed.2018.3763)] [Medline: [30128552](https://pubmed.ncbi.nlm.nih.gov/30128552/)]
41. National survey on drug use and health (NSDUH): population data. Substance Abuse and Mental Health Association (SAMHSA). URL: <https://www.datafiles.samhsa.gov/dataset/national-survey-drug-use-and-health-2020-nsduh-2020-ds0001> [accessed 2024-04-29]
42. Opioids. Johns Hopkins Medicine. URL: <https://www.hopkinsmedicine.org/health/treatment-tests-and-therapies/opioids#:~:text=Opioids%20are%20a%20class%20of,pain%20medicine%20and%20illegal%20drugs> [accessed 2024-04-29]
43. Nalini M, Khoshnia M, Kamangar F, Sharafkhah M, Poustchi H, Pourshams A, et al. Joint effect of diabetes and opiate use on all-cause and cause-specific mortality: the Golestan cohort study. *Int J Epidemiol* 2021 Mar 03;50(1):314-324 [FREE Full text] [doi: [10.1093/ije/dyaa126](https://doi.org/10.1093/ije/dyaa126)] [Medline: [32810213](https://pubmed.ncbi.nlm.nih.gov/32810213/)]

44. Vozoris NT, Wang X, Fischer HD, Bell CM, O'Donnell DE, Austin PC, et al. Incident opioid drug use and adverse respiratory outcomes among older adults with COPD. *Eur Respir J* 2016 Sep 13;48(3):683-693 [FREE Full text] [doi: [10.1183/13993003.01967-2015](https://doi.org/10.1183/13993003.01967-2015)] [Medline: [27418553](https://pubmed.ncbi.nlm.nih.gov/27418553/)]
45. Rogal S, Youk A, Agbalajobi O, Zhang H, Gellad W, Fine MJ, et al. Medication treatment of active opioid use disorder in veterans with cirrhosis. *Am J Gastroenterol* 2021 Jul 01;116(7):1406-1413 [FREE Full text] [doi: [10.14309/ajg.0000000000001228](https://doi.org/10.14309/ajg.0000000000001228)] [Medline: [33811202](https://pubmed.ncbi.nlm.nih.gov/33811202/)]
46. Rosenthal ES, Silk R, Mathur P, Gross C, Eyasu R, Nussdorf L, et al. Concurrent initiation of Hepatitis C and Opioid use disorder treatment in people who inject drugs. *Clin Infect Dis* 2020 Oct 23;71(7):1715-1722 [FREE Full text] [doi: [10.1093/cid/ciaa105](https://doi.org/10.1093/cid/ciaa105)] [Medline: [32009165](https://pubmed.ncbi.nlm.nih.gov/32009165/)]
47. Johansson ED, Nunez M. Acute Hepatitis B surge: opioid epidemic implication and management challenges. *Open Forum Infect Dis* 2020 Jun;7(6):ofaa190 [FREE Full text] [doi: [10.1093/ofid/ofaa190](https://doi.org/10.1093/ofid/ofaa190)] [Medline: [32550238](https://pubmed.ncbi.nlm.nih.gov/32550238/)]
48. Owsiany MT, Hawley CE, Triantafylidis LK, Paik JM. Opioid management in older adults with chronic kidney disease: a review. *Am J Med* 2019 Dec;132(12):1386-1393 [FREE Full text] [doi: [10.1016/j.amjmed.2019.06.014](https://doi.org/10.1016/j.amjmed.2019.06.014)] [Medline: [31295441](https://pubmed.ncbi.nlm.nih.gov/31295441/)]
49. Naik R, Goodrich G, Al-Shaikhly T, Joks R. Prevalence of long term opioid use in patients with asthma and allergic rhinitis. *J Allergy Clin Immunol* 2018 Feb;141(2):AB218. [doi: [10.1016/j.jaci.2017.12.690](https://doi.org/10.1016/j.jaci.2017.12.690)]
50. Cunningham CO. Opioids and HIV infection: from pain management to addiction treatment. *Top Antivir Med* 2018 Apr;25(4):143-146 [FREE Full text] [Medline: [29689538](https://pubmed.ncbi.nlm.nih.gov/29689538/)]
51. Ganguly A, Michael M, Goschin S, Harris K, McFarland DC. Cancer pain and Opioid use disorder. *Oncology (Williston Park)* 2022 Sep 07;36(9):535-541 [FREE Full text] [doi: [10.46883/2022.25920973](https://doi.org/10.46883/2022.25920973)] [Medline: [36107782](https://pubmed.ncbi.nlm.nih.gov/36107782/)]
52. Tumenta T, Ugwendum DF, Chobufo MD, Mungu EB, Kogan I, Olupona T. Prevalence and trends of Opioid use in patients with depression in the United States. *Cureus* 2021 May 28;13(5):e15309 [FREE Full text] [doi: [10.7759/cureus.15309](https://doi.org/10.7759/cureus.15309)] [Medline: [34221762](https://pubmed.ncbi.nlm.nih.gov/34221762/)]
53. Stokes A, Lundberg DJ, Hempstead K, Berry KM, Baker JF, Preston SH. Obesity and incident prescription Opioid use in the U.S., 2000-2015. *Am J Prev Med* 2020 Jun;58(6):766-775 [FREE Full text] [doi: [10.1016/j.amepre.2019.12.018](https://doi.org/10.1016/j.amepre.2019.12.018)] [Medline: [32229057](https://pubmed.ncbi.nlm.nih.gov/32229057/)]
54. Stokes A, Lundberg DJ, Sheridan B, Hempstead K, Morone NE, Lasser KE, et al. Association of obesity with prescription opioids for painful conditions in patients seeking primary care in the US. *JAMA Netw Open* 2020 Apr 01;3(4):e202012 [FREE Full text] [doi: [10.1001/jamanetworkopen.2020.2012](https://doi.org/10.1001/jamanetworkopen.2020.2012)] [Medline: [32239222](https://pubmed.ncbi.nlm.nih.gov/32239222/)]
55. Opioid use disorder. Johns Hopkins Medicine. URL: <https://www.hopkinsmedicine.org/health/conditions-and-diseases/opioid-use-disorder> [accessed 2024-04-29]
56. Hoffman JI. Logistic regression. In: Hoffman JI, editor. *Biostatistics for Medical and Biomedical Practitioners*. Cambridge, MA: Academic Press; 2015:601-611.
57. Cervantes J, Garcia-Lamont F, Rodríguez-Mazahua L, Lopez A. A comprehensive survey on support vector machine classification: applications, challenges and trends. *Neurocomputing* 2020 Sep;408:189-215 [FREE Full text] [doi: [10.1016/j.neucom.2019.10.118](https://doi.org/10.1016/j.neucom.2019.10.118)]
58. Penm J, Chaar B, Moles R, Penm J. Predicting ASX health care stock index movements after the recent financial crisis using patterned neural networks. In: Wehn CS, Hoppe C, Gregoriou GN, editors. *Rethinking Valuation and Pricing Models: Lessons Learned from the Crisis and Future Challenges*. Cambridge, MA: Academic Press; 2023:599-610.
59. Abadi M, Barham P, Chen J, Chen Z, Davis A, Dean J, et al. TensorFlow: a system for large-scale machine learning. In: *Proceedings of the 12th USENIX conference on Operating Systems Design and Implementatio*. 2016 Presented at: OSDI '16; November 2-4, 2016; Savannah, GA p. 265-283 URL: <https://dl.acm.org/doi/10.5555/3026877.3026899>
60. Ruder S. An overview of gradient descent optimization algorithms. arXiv Preprint posted online September 15, 2016 [FREE Full text] [doi: [10.1017/9781108699211.008](https://doi.org/10.1017/9781108699211.008)]
61. Kingma DP, Ba J. Adam: a method for stochastic optimization. In: *Proceedings of the 2015 International Conference on Learning Representations*. 2015 Presented at: ICLR '15; May 7-9, 2015; San Diego, CA p. 1-7 URL: <https://dblp.org/rec/journals/corr/KingmaB14.html>
62. Hardt M, Price E, Srebro N. Equality of opportunity in supervised learning. In: *Proceedings of the 30th International Conference on Neural Information Processing Systems*. 2016 Presented at: ICNIPS '16; December 5-10, 2016; Barcelona, Spain p. 3323 URL: <https://dl.acm.org/doi/10.5555/3157382.3157469>
63. Sivarajkumar S, Huang Y, Wang Y. Fair patient model: mitigating bias in the patient representation learned from the electronic health records. *J Biomed Inform* 2023 Dec;148:104544. [doi: [10.1016/j.jbi.2023.104544](https://doi.org/10.1016/j.jbi.2023.104544)] [Medline: [37995843](https://pubmed.ncbi.nlm.nih.gov/37995843/)]
64. Li F, Wu P, Ong HH, Peterson JF, Wei WQ, Zhao J. Evaluating and mitigating bias in machine learning models for cardiovascular disease prediction. *J Biomed Inform* 2023 Feb;138:104294 [FREE Full text] [doi: [10.1016/j.jbi.2023.104294](https://doi.org/10.1016/j.jbi.2023.104294)] [Medline: [36706849](https://pubmed.ncbi.nlm.nih.gov/36706849/)]
65. Working together to reduce black maternal mortality. Centers for Disease Control and Prevention. URL: <https://www.cdc.gov/healthequity/features/maternal-mortality/index.html#:~:text=Racial%20Disparities%20Exist,structural%20racism%2C%20and%20implicit%20bias> [accessed 2024-04-29]

66. Juhn YJ, Ryu E, Wi CI, King KS, Malik M, Romero-Brufau S, et al. Assessing socioeconomic bias in machine learning algorithms in health care: a case study of the HOUSES index. *J Am Med Inform Assoc* 2022 Jun 14;29(7):1142-1151 [FREE Full text] [doi: [10.1093/jamia/ocac052](https://doi.org/10.1093/jamia/ocac052)] [Medline: [35396996](https://pubmed.ncbi.nlm.nih.gov/35396996/)]
67. Chow EA, Foster H, Gonzalez V, McIver L. The disparate impact of diabetes on racial/ethnic minority populations. *Clin Diabetes* 2012 Jul 16;30(3):130-133. [doi: [10.2337/diaclin.30.3.130](https://doi.org/10.2337/diaclin.30.3.130)]
68. Depression in women: understanding the gender gap. Mayo Clinic. 2019. URL: <https://www.mayoclinic.org/diseases-conditions/depression/in-depth/depression/art-20047725> [accessed 2024-04-29]
69. Hicks SA, Strümke I, Thambawita V, Hammou M, Riegler MA, Halvorsen P, et al. On evaluation metrics for medical applications of artificial intelligence. *Sci Rep* 2022 Apr 08;12(1):5979 [FREE Full text] [doi: [10.1038/s41598-022-09954-8](https://doi.org/10.1038/s41598-022-09954-8)] [Medline: [35395867](https://pubmed.ncbi.nlm.nih.gov/35395867/)]
70. Burlew K, McCuistian C, Szapocznik J. Racial/ethnic equity in substance use treatment research: the way forward. *Addict Sci Clin Pract* 2021 Aug 05;16(1):50 [FREE Full text] [doi: [10.1186/s13722-021-00256-4](https://doi.org/10.1186/s13722-021-00256-4)] [Medline: [34353373](https://pubmed.ncbi.nlm.nih.gov/34353373/)]
71. Schepis TS, Wastila L, Ammerman B, McCabe VV, McCabe SE. Prescription opioid misuse motives in US older adults. *Pain Med* 2020 Oct 01;21(10):2237-2243 [FREE Full text] [doi: [10.1093/pm/pnz304](https://doi.org/10.1093/pm/pnz304)] [Medline: [31816076](https://pubmed.ncbi.nlm.nih.gov/31816076/)]

Abbreviations

AUC: area under the curve
DL: deep learning
EO: equality of odds
LR: logistic regression
ML: machine learning
NN: neural network
NN-SGD: neural network model using stochastic gradient descent
NSDUH: National Survey on Drug Use and Health
OD: opioid overdose
OOD: opioid use disorder
PR: precision-recall
RBF: radial basis function
ROC: receiver operating characteristic
SGD: stochastic gradient descent
SVM: support vector machine
WMV: weighted majority voting

Edited by K El Emam, B Malin; submitted 25.12.23; peer-reviewed by S Matsuda, D Harris; comments to author 11.05.24; revised version received 22.06.24; accepted 29.06.24; published 20.08.24.

Please cite as:

Yaseliani M, Noor-E-Alam M, Hasan MM

Mitigating Sociodemographic Bias in Opioid Use Disorder Prediction: Fairness-Aware Machine Learning Framework

JMIR AI 2024;3:e55820

URL: <https://ai.jmir.org/2024/1/e55820>

doi: [10.2196/55820](https://doi.org/10.2196/55820)

PMID:

©Mohammad Yaseliani, Md Noor-E-Alam, Md Mahmudul Hasan. Originally published in JMIR AI (<https://ai.jmir.org>), 20.08.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR AI, is properly cited. The complete bibliographic information, a link to the original publication on <https://www.ai.jmir.org/>, as well as this copyright and license information must be included.

Original Paper

Traditional Machine Learning, Deep Learning, and BERT (Large Language Model) Approaches for Predicting Hospitalizations From Nurse Triage Notes: Comparative Evaluation of Resource Management

Dhavalkumar Patel^{1*}, MSc; Prem Timsina¹, PhD; Larisa Gorenstein², MD; Benjamin S Glicksberg³, PhD; Ganesh Raut¹, MSc; Satya Narayan Cheetirala¹, MSc; Fabio Santana¹, BSc; Jules Tamegue¹, BSc; Arash Kia¹, MD; Eyal Zimlichman^{4,5}, MD; Matthew A Levin^{6,7,8}, MD; Robert Freeman¹, PhD; Eyal Klang^{3*}, MD

¹Institute for Healthcare Delivery Science, Icahn School of Medicine at Mount Sinai, New York, NY, United States

²Division of Diagnostic Imaging, Sheba Medical Center, Tel-Aviv University, Tel Aviv, Israel

³Division of Data-Driven and Digital Medicine (D3M), Icahn School of Medicine at Mount Sinai, New York, NY, United States

⁴Hospital Management, Sheba Medical Center, Tel-Aviv University, Tel Aviv, Israel

⁵ARC Innovation Center, Sheba Medical Center, Tel-Aviv University, Tel Aviv, Israel

⁶Department of Anesthesiology, Perioperative and Pain Medicine, Icahn School of Medicine at Mount Sinai, New York, NY, United States

⁷Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY, United States

⁸Windreich Department of Artificial Intelligence and Human Health, Icahn School of Medicine at Mount Sinai, New York, NY, United States

*these authors contributed equally

Corresponding Author:

Dhavalkumar Patel, MSc

Institute for Healthcare Delivery Science

Icahn School of Medicine at Mount Sinai

2nd Floor

150 East 42nd Street

New York, NY, 10017

United States

Phone: 1 (212) 523 5555

Email: pateldhaval021@hotmail.com

Abstract

Background: Predicting hospitalization from nurse triage notes has the potential to augment care. However, there needs to be careful considerations for which models to choose for this goal. Specifically, health systems will have varying degrees of computational infrastructure available and budget constraints.

Objective: To this end, we compared the performance of the deep learning, Bidirectional Encoder Representations from Transformers (BERT)-based model, Bio-Clinical-BERT, with a bag-of-words (BOW) logistic regression (LR) model incorporating term frequency-inverse document frequency (TF-IDF). These choices represent different levels of computational requirements.

Methods: A retrospective analysis was conducted using data from 1,391,988 patients who visited emergency departments in the Mount Sinai Health System spanning from 2017 to 2022. The models were trained on 4 hospitals' data and externally validated on a fifth hospital's data.

Results: The Bio-Clinical-BERT model achieved higher areas under the receiver operating characteristic curve (0.82, 0.84, and 0.85) compared to the BOW-LR-TF-IDF model (0.81, 0.83, and 0.84) across training sets of 10,000; 100,000; and ~1,000,000 patients, respectively. Notably, both models proved effective at using triage notes for prediction, despite the modest performance gap.

Conclusions: Our findings suggest that simpler machine learning models such as BOW-LR-TF-IDF could serve adequately in resource-limited settings. Given the potential implications for patient care and hospital resource management, further exploration of alternative models and techniques is warranted to enhance predictive performance in this critical domain.

International Registered Report Identifier (IRRID): RR2-10.1101/2023.08.07.23293699

KEYWORDS

Bio-Clinical-BERT; term frequency–inverse document frequency; TF-IDF; health informatics; patient care; hospital resource management; care; resource management; management; language model; machine learning; hospitalization; deep learning; logistic regression; retrospective analysis; training; large language model

Introduction

Efficient and effective patient triage within the emergency department (ED) plays a pivotal role in enhancing treatment outcomes and optimizing care delivery [1-3]. This process involves rapidly identifying patients who require immediate hospitalization upon their arrival. One of the resources for making these predictions are nurse triage notes, which provide a wealth of in-depth information about the patient's condition at presentation [4,5].

In the field of health care, machine learning has opened up new avenues for potential improvement in such complex classification tasks, thereby augmenting clinical decision-making processes [6,7]. The recent developments in deep learning and natural language processing (NLP) techniques have further broadened this potential, bringing a new realm of possibilities for enhancing medical decision-making capabilities.

Among these advanced algorithms is the Bidirectional Encoder Representations from Transformers (BERT) model [8]. BERT has shown excellent performance in numerous NLP tasks [9] and has inspired the development of more specialized versions tailored to particular fields, such as the Bio-Clinical-BERT model, which was designed to cater to the biomedical field [10].

The focus of this study is to delve into the potential of a fine-tuned Bio-Clinical-BERT model and compare it against a simpler, robust, and more traditional approach, mainly, the bag-of-words (BOW) logistic regression (LR) model complemented by the term frequency–inverse document frequency (TF-IDF) method. We also evaluated other approaches including the extreme gradient boosting (XGBoost) classifier and Word-2-Vec (W2V) embedding with bidirectional long short-term memory (Bi-LSTM) network. The primary objective of our research is to gauge the efficacy of these 2 methods in predicting hospital admissions using nurse triage notes.

While it is true that Bio-Clinical-BERT could potentially offer improved accuracy in its predictions, it should be noted that it also requires a substantial investment in terms of computational resources. It necessitates the use of specialized hardware and demands a certain level of software expertise to operate effectively. On the other hand, the LR model paired with the TF-IDF method is more resource efficient and enjoys wide acceptance in the field of text classification due to its simplicity and effectiveness.

We hypothesized that the Bio-Clinical-BERT model may surpass the performance of the BOW-LR model combined with the TF-IDF approach in the task of predicting triage outcomes. However, we also speculated that the incremental gains in performance might not necessarily justify the additional

demands imposed by the large deep learning model in terms of computational resources and technical know-how. To test this hypothesis, we have undertaken an extensive study using over 1 million nurse triage notes collected from a large health system.

The fundamental contribution of this paper is a comparison between these techniques for predicting hospital admission, which reflect different levels of computational requirements and cost implications. Our comparison not only looks at the accuracy of these models but also weighs the trade-offs between predictive accuracy and computational efficiency, a consideration that is often overlooked but is of prime importance in real-world settings when implementing models. Specifically, health systems may be able to use insights from this study to make informed decisions on which methodology may be right for their circumstances, with a clearer understanding of the limitations of each. Our aim is to equip health care practitioners, researchers, and decision makers with insights that could potentially aid in enhancing hospital resource management and improve the quality of patient care.

Methods

Data Sources and Study Design

For the construction and testing of our models, we used an extensive dataset from the Mount Sinai Health System (MSHS). This is a diverse health care provider based in New York City. In this study, the dataset included ED records spanning a 5-year period from 2017 to 2022. This dataset was gathered from 5 different MSHS hospitals, covering a broad range of population groups and diverse urban health settings.

These 5 participating hospitals provided a rich source of data for our study, representing different communities in New York City. The hospitals include Mount Sinai Hospital, a health care institution located in East Harlem, Manhattan; Mount Sinai Morningside, situated in Morningside Heights, Manhattan; Mount Sinai West, operating in Midtown West, Manhattan; Mount Sinai Brooklyn, a community-focused health facility located in Midwood, Brooklyn; and Mount Sinai Queens (MSQ), based in Astoria, Queens. The dataset used for our study was compiled using the Epic Electronic Health Records software, a tool that aids in efficient data collection, management, and analysis. The dataset was made available by the diligent work of the Mount Sinai Hospital Clinical Data Science team.

Model Development and Evaluation

In the development and testing of our models, we leveraged data from 4 hospitals for training, validation, and hyperparameter tuning processes. We elected to use a distinct dataset from MSQ for external testing to ensure our model's generalizability.

The internal training and validation cohort underwent a procedure involving 5-fold cross-validation. Each fold contained 10,000 records, which were used for hyperparameter tuning. For the external dataset, we experimented with training sets of varying sizes: 10,000; 100,000; and roughly 1,000,000 patients, which represent the complete 4-hospital cohort. Subsequently, testing was carried out on 20% of these cohorts' sizes, taken from the MSQ hospital cohort.

Our study involved several models: Bio-Clinical-BERT and BOW-LR models using TF-IDF features. For further subanalyses using different machine and deep learning models, we also evaluated XGBoost with BOW and Bi-LSTM with a W2V pretrained embedding layer derived from bioclinical data (BioWordVec_PubMed_MIMICIII_d200).

As a final subanalysis experiment, for the BERT model, we also experimented with up-sampling of the minority class to ensure balanced data representation, enhancing the stability and accuracy of our model predictions.

These models were used to predict hospitalization outcomes from nurse triage notes. For Bio-Clinical-BERT, we adhered to text preprocessing and tokenization guidelines as outlined on the Hugging Face website [11].

For BOW-XGBoost, we evaluated 3 different numbers of estimators. Other XGBoost hyperparameters were set to default values, including a learning rate of 0.3, maximum depth of 6, and minimum child weight of 1.

For W2V-Bi-LSTM, the network is comprised of a Bi-LSTM layer (256 hidden units), preceded by a pretrained embedding 200-dimensions W2V layer, with a fully connected layer followed by a sigmoid activation function.

Further details on hyperparameter selection are elucidated in the *Hyperparameter Tuning Results* section. For BOW-LR-TF-IDF, we followed a similar methodology outlined in our previous publication [12], covering both text preprocessing and hyperparameter selection processes.

BERT is a model designed for NLP tasks. It learns from the context of both preceding and following words, making it "bidirectional." This model is pretrained on large corpora and can be fine-tuned for specific tasks.

The BOW model is a simple technique in NLP. It represents text data by counting the frequency of each word, disregarding the order in which they appear. Each unique word forms a feature, and the frequency of the word represents the value of that feature. However, this method can overlook context and semantics due to its simplicity.

TF-IDF is a numerical statistic that reflects how important a word is to a document in a collection. It is a combination of 2 metrics: *term frequency*, which is the number of times a word appears in a document, and *inverse document frequency*, which diminishes the weight of common words and amplifies the weight of rare words across the entire dataset. This helps in reducing the impact of frequently used words and highlights more meaningful terms.

XGBoost is an advanced gradient boosting framework known for its efficiency and performance in structured data classification and regression. It builds multiple decision trees sequentially to correct previous errors, excelling in handling diverse data types and preventing overfitting.

Bi-LSTM is an artificial neural network that processes data in both directions to capture past and future context. This enhances its sequence understanding, making it suitable for text classification, sentiment analysis, and machine translation.

Study Population

The demographic for this study included adult patients aged 18 years and older. These were patients who made ED visits within the specified 5-year period from 2017 to 2022 across the 5 participating MSHS hospitals.

Outcome Definition

The primary outcome for our study was to ascertain our models' effectiveness in predicting hospitalization. This prediction was based on 2 main types of data: tabular electronic health records and nurse triage notes.

Model Evaluation and Comparison

To assess the performance of our models, we used various metrics such as area under the receiver operating characteristic curve (AUC), sensitivity, specificity, and precision. These metrics allowed us to thoroughly evaluate the Bio-Clinical-BERT [10] and BOW-LR models with TF-IDF features, as well as compare their capabilities in predicting hospitalization from nurse triage notes.

Ethical Considerations

This study, being retrospective in nature, was reviewed and approved by an ethical institutional review board committee from the MSHS (protocol: STUDY-18-00573). The institutional review board committee deemed that due to the retrospective nature of the study, the requirement for informed consent was waived.

Statistical Analysis

Our statistical analyses were conducted using Python (version 3.9.12; Python Software Foundation). We presented continuous variables as median (IQR) and categorical variables as percentages for better interpretability. To identify words linked to hospital admission within nurse triage notes, we calculated the odds ratio (OR) and mutual information (MI) [12]. Statistical tests such as the chi-square test and 2-tailed Student *t* test were used for comparing categorical and continuous variables, respectively. A *P* value <.05 was considered statistically significant. For evaluating our models, receiver operating characteristic (ROC) curves were plotted, and metrics including AUC, sensitivity (recall), specificity, and positive predictive value (precision) were derived.

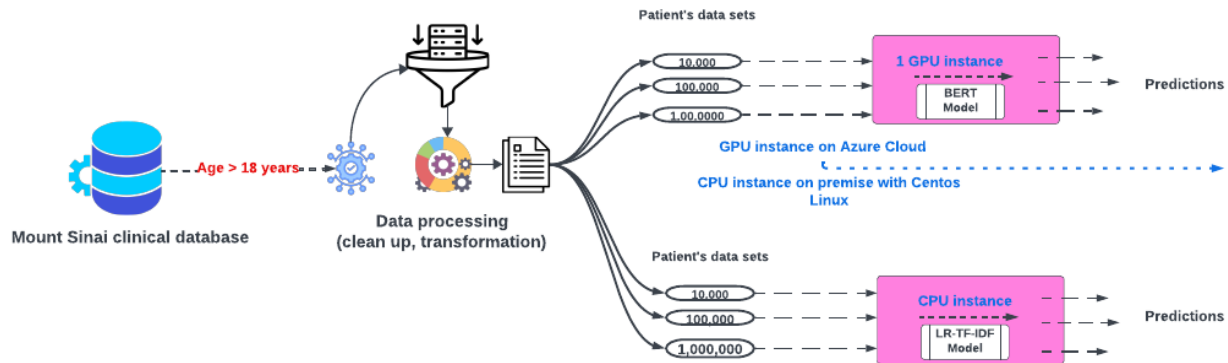
Technical Architecture

The technical experiments involved in this study were conducted within a controlled hospital infrastructure that used an On-Premises Centos Linux environment in conjunction with Azure Cloud infrastructure. For the BOW-TF-IDF experiments,

we elected to use the Centos Linux OS. In contrast, the BERT experiment was conducted using a Standard_NC6 GPU instance on Azure Cloud. This instance came with one 16-GB GPU and 6 vCPUs and incurred a cost of approximately US \$80 during

the training phase. Figure 1 offers a detailed depiction of the fundamental technical architecture used for training the BERT and LR-TF-IDF models, across multiple patient datasets.

Figure 1. Process flow of multiple patient datasets passing through 2 different models with GPU and non-GPU instances. BERT: Bidirectional Encoder Representations from Transformers; LR: logistic regression; TF-IDF: term frequency–inverse document frequency.



Results

Patient Population and Data

Our study incorporated data from 1,745,199 patients drawn from the MSHS. Upon the exclusion of patients aged <18 years,

we had 1,391,988 participants in the study. These patients visited the ED between 2017 and 2022. Table 1 presents a summary of the patient characteristics.

The median number of words per triage note was 19.0 (IQR 12.0-31.0). Top 10 words associated with the highest MI score regarding hospital admission are outlined in Table 2.

Table 1. Demographic distribution in the study.

Demographics	All patients (includes MSH ^a , MSM ^b , MSW ^c , MSB ^d , and MSQ ^e ; N=1,391,988)	4 hospitals (MSH, MSM, MSW, and MSB; n=1,110,272)	MSQ (n=281,716)	P value
Age (years), median (IQR)	47.0 (31.0-75)	48.0 (32.0-75.0)	45.0 (30.0-75.0)	<.001
Sex, n (%)				<.001
Female	727,363 (52.3)	586,224 (52.8)	141,140 (50.1)	
Male	664,625 (47.7)	524,048 (47.2)	140,576 (49.9)	
Race, n (%)				<.001
Black	428,594 (30.79)	382,898 (34.5)	45,696 (16.22)	
White	343,079 (24.65)	265,457 (23.92)	77,622 (27.56)	
Other	620,315 (44.56)	461,917 (41.58)	158,398 (56.22)	

^aMSH: Mount Sinai Hospital.

^bMSM: Mount Sinai Morningside.

^cMSW: Mount Sinai West.

^dMSB: Mount Sinai Brooklyn.

^eMSQ: Mount Sinai Queens.

Table 2. Odds ratios (OR) and mutual information (MI) values for words linked to admission to hospital wards, sorted by highest MI values.

Word	OR for admission	MI for admission	P value
Sent	3.6	16.4	<.001
Pt ^a	1.6	15.8	<.001
Per	2.3	15	<.001
Of	1.3	12.7	<.001
Home	2.2	11.5	<.001
EMS ^b	2.2	10.8	<.001
Weakness	3.6	10.8	<.001
Chest	1.4	8.9	<.001
SOB ^c	2.1	8.8	<.001
BIBA ^d	2.1	7.9	<.001

^aPt: patient.

^bEMS: emergency medical services.

^cSOB: shortness of breath.

^dBIBA: brought in by ambulance.

Hyperparameter Tuning Results

A hyperparameter tuning process was performed. The best hyperparameters were identified for each model based on their performance during the 5-fold cross-validation on the training validation set. The results of the BERT hyperparameter tuning process can be found in [Table 3](#).

The results of the W2V-LSTM model hyperparameter tuning are presented in [Table 4](#).

The results of XGBoost hyperparameter tuning are presented in [Table 5](#).

Table 3. BERT^a hyperparameter tuning in the internal training and validation cohorts using 5-fold experiments.

Batch size	Max length	Learning rate	Epoch	Value, mean (SD)
64	— ^b	2×10^{-5}	—	0.78 (0.01)
128	—	2×10^{-5}	—	0.80 (0.01)
128	128	2×10^{-5}	3	0.80 (0.01)
256	64	2×10^{-5}	—	0.79 (0.01)
64	—	3×10^{-5}	—	0.79 (0.01)
128	—	3×10^{-5}	—	0.79 (0.01)
128	128	3×10^{-5}	3	0.78 (0.01)
256	64	3×10^{-5}	—	0.78 (0.01)
64	—	5×10^{-5}	—	0.79 (0.01)
128	—	5×10^{-5}	—	0.80 (0.01)
128	128	5×10^{-5}	3	0.79 (0.01)
256	64	5×10^{-5}	—	0.79 (0.01)

^aBERT: Bidirectional Encoder Representations from Transformers.

^bNot applicable.

Table 4. W2V^a-LSTM^b hyperparameter tuning in the internal training and validation cohorts using 5-fold experiments.

Batch size	Learning rate	Epochs	AUC ^c , mean
16	10 ⁻³	5	0.768
16	10 ⁻³	10	0.795
16	10 ⁻³	15	0.765
16	10 ⁻⁴	5	0.750
16	10 ⁻⁴	10	0.781
16	10 ⁻⁴	15	0.798
32	10 ⁻³	5	0.797
32	10 ⁻³	10	0.797
32	10 ⁻³	15	0.777
32	10 ⁻⁴	5	0.756
32	10 ⁻⁴	10	0.728
32	10 ⁻⁴	15	0.748
64	10 ⁻³	5	0.661
64	10 ⁻³	10	0.806
64	10 ⁻³	15	0.795
64	10 ⁻⁴	5	0.693
64	10 ⁻⁴	10	0.767
64	10 ⁻⁴	15	0.775

^aW2V: Word-2-Vec.

^bLSTM: long short-term memory.

^cAUC: area under the receiver operating characteristic curve.

Table 5. XGBoost^a hyperparameter tuning in the internal training and validation cohorts using 5-fold experiments.

Trees	Value
100	0.80 (0.01)
200	0.81 (0.01)
1000	0.80 (0.01)

^aXGBoost: extreme gradient boosting.

Model Performance

After training the Bio-Clinical-BERT and LR-TF-IDF models on the 4 hospitals' data, we evaluated their performance on the held-out test data from MSQ. The AUC values were calculated for each model. The Bio-Clinical-BERT model achieved AUCs of 0.82, 0.84, 0.85, while the LR-TF-IDF model had AUCs of 0.81, 0.83, 0.84 for training on 10,000; 100,000; and ~1,000,000 patients, respectively.

Figure 2 shows the ROC and AUC comparisons between the 2 models. The Bio-Clinical-BERT model consistently outperformed the LR-TF-IDF model in terms of AUC across the different training set sizes (10,000; 100,000; and ~1,000,000 patients), albeit by a small margin.

In addition to the AUC comparisons, we also calculated other performance metrics, such as sensitivity, specificity, and precision, for both models (Table 6 and Table 7).

Figure 2. Receiver operating characteristic curves (ROC) of the 2 models tested on increasing training sample sizes. AUC: area under the receiver operating characteristic curve; BERT: Bidirectional Encoder Representations from Transformers; LR: logistic regression; MSQ: Mount Sinai Queens; TF-IDF: term frequency–inverse document frequency.

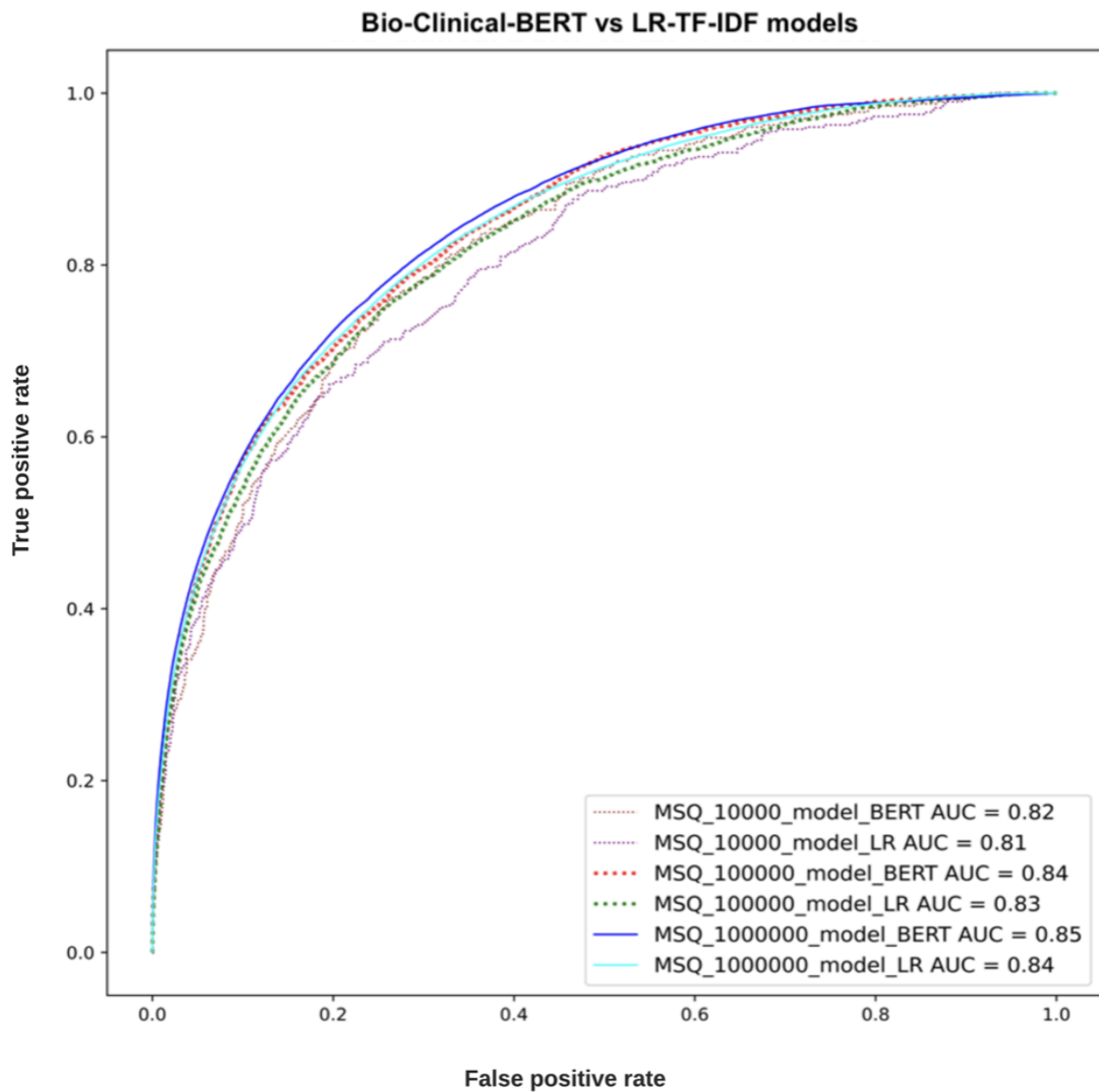


Table 6. Metrics for the training and testing (external) cohorts for the Bio-Clinical-BERT^a model.

Training data size	AUC ^b score	Sensitivity	Specificity	Precision	F_1 -score
10,000	0.82	0.76	0.74	0.36	0.49
100,000	0.84	0.74	0.77	0.39	0.51
1,000,000	0.85	0.39	0.96	0.67	0.50

^aBERT: Bidirectional Encoder Representations from Transformers.

^bAUC: area under the receiver operating characteristic curve.

Table 7. Metrics for the training and testing (external) cohorts for the LR^a-TF-IDF^b model.

Training Data Size	AUC ^c score	Sensitivity	Specificity	Precision	F_1 -score
10,000	0.81	0.66	0.80	0.40	0.50
100,000	0.83	0.75	0.74	0.37	0.50
1,000,000	0.84	0.71	0.80	0.42	0.53

^aLR: logistic regression.

^bTF-IDF: term frequency–inverse document frequency.

^cAUC: area under the receiver operating characteristic curve.

The metrics for the XGBoost and W2V-Bi-LSTM models are presented in [Tables 8](#) and [9](#). The probability cutoff values for these metrics were calculated using the Youden index. These

results further demonstrated the superior performance of the Bio-Clinical-BERT model compared to the LR-TF-IDF model.

Further subanalysis for the BERT cohort using up-sampling of the minority class is presented in [Table 10](#).

Table 8. Metrics for the training and testing (external) cohorts for the W2V^a-Bi-LSTM^b model.

Training data size	AUC ^c score	Sensitivity	Specificity	Precision	F_1 -score
10,000	0.78	0.32	0.95	0.59	0.41
100,000	0.81	0.42	0.92	0.52	0.46
1,000,000	0.84	0.46	0.94	0.62	0.52

^cW2V: Word-2-Vec.

^cBi-LSTM: bidirectional long short-term memory.

^cAUC: area under the receiver operating characteristic curve.

Table 9. Metrics for the training and testing (external) cohorts for the XGBoost^a model.

Training data size	AUC ^b score	Sensitivity	Specificity	Precision	F_1 -score
10,000	0.76	0.21	0.97	0.69	0.33
100,000	0.81	0.27	0.98	0.73	0.39
1,000,000	0.82	0.33	0.97	0.69	0.45

^aXGBoost: extreme gradient boosting.

^bAUC: area under the receiver operating characteristic curve.

Table 10. Metrics for the training and testing (external) cohorts for BERT^a with up-sampling of the minority class.

Training data size	AUC ^b score	Sensitivity	Specificity	Precision	F_1 -score
10,000	0.79	0.81	0.58	0.30	0.43
100,000	0.84	0.81	0.68	0.34	0.48
1,000,000	0.85	0.75	0.78	0.41	0.54

^aBERT: Bidirectional Encoder Representations from Transformers.

^bAUC: area under the receiver operating characteristic curve.

Discussion

In this study, we compared the performance of several predictive models, including Bio-Clinical-BERT and LR-TF-IDF, in predicting hospitalizations based on nurse triage notes. The findings of our study suggest that while Bio-Clinical-BERT does marginally outperform LR-TF-IDF in this predictive task, the difference in their performance is relatively minor.

Such results echo the findings of previous studies in the field, which have often found BERT-based models to have a slight

edge over more traditional methods such as LR-TF-IDF in various NLP tasks [13,14]. However, the marginal difference observed in our study suggests that, given certain limitations such as constraints on hardware, software expertise, or budget, hospitals might lean toward simpler methods. The rationale behind such a choice would lie in the ease of implementing these simpler methods, as well as their relatively less demanding computational requirements.

The comparison of different models in the biomedical domain has been the focus of numerous previous studies. For instance,

Chen et al [15] conducted an assessment of transformer-based ChatGPT models in tasks such as reasoning and classification. Their study found that fine-tuning remained the most effective approach for 2 central NLP tasks. However, it is interesting to note that the basic BOW model managed to deliver comparable results to the more complex language model prompting. It should be noted that the creation of effective prompts required a substantial resource investment.

In another study, Xavier and Chen [16] compared 3 different model types for a multiclass text classification task, which involved the assignment of protocols for abdominal imaging computed tomography scans. These models spanned a range from conventional machine learning and deep learning to automated machine learning builder workflows. While the automated machine learning builder boasted the best performance with an F_1 -score of 0.85 on an unbalanced dataset, the tree ensemble machine learning algorithm was superior on a balanced dataset, delivering an F_1 -score of 0.80.

A further study delved into the evaluation of machine learning multiclass classification algorithms' performance in classifying proximal humeral fractures using radiology text data [17]. Several statistical machine learning algorithms were performed, with a BERT model showcasing the best accuracy of 61%. In another relevant study conducted by Ji et al [18], various models pretrained with BERT were compared for medical code assignment based on clinical notes. Interestingly, it was found that simpler artificial neural networks could sometimes outperform BERT in certain scenarios. This study, among others, offers further support to our recommendation for hospitals with limited resources to consider simpler, less resource-demanding methods for achieving comparable predictive performance.

In the specific task of predicting hospitalization, both methods in our study effectively leveraged the rich information found within nurse triage notes. This finding aligns with those from other studies [19-21]. For instance, a study by Zhang et al [19] that evaluated LR and neural network modeling approaches in predicting hospital admission or transfer after initial ED triage presentation found that the patient's free-text data regarding referral improved overall predictive accuracy. Similarly, Raita et al [20] used machine learning models to predict ED outcomes and demonstrated superior performance in predicting hospitalization.

The results of our study carry practical implications for health care organizations. The ability to predict hospitalization from nurse triage notes could lead to improvements in patient care by facilitating efficient resource allocation, optimizing bed management, and improving patient flow.

The choice between the use of Bio-Clinical-BERT and simpler methods, such as LR-TF-IDF, should be influenced by the specific context of the organization, including factors such as available computational resources, software expertise, and desired model performance.

Our study is not without limitations. For instance, the data used for our study are specific to MSHS hospitals, which might not be representative of other health care systems, potentially limiting the generalizability of our findings. Despite using multisite data, representing the diverse New York City population, and an external validation site for our final analysis, we acknowledge the need for further studies with more diverse datasets, including those that are open source such as the Medical Information Mart for Intensive Care (MIMIC) dataset. We also recognize that we did not explore the potential of combining both methods and other potential techniques that could enhance these models' performance. The BOW technique by nature does not consider context, which could have hindered performance. There is the possibility that more advanced deep learning models could have achieved a bigger difference in AUC performance compared to the shallow model. Moreover, the field of NLP is advancing fast, and some methodologies were not explored. Also, our study focused on comparative analysis using the Youden index, which may have caused several metrics to be lower than previous publications, such as the F_1 -score. Despite this, the models demonstrated high specificity, suggesting potential for clinical use. Further exploration of thresholding methods is necessary to enhance model applicability and performance in real-world settings.

Future research could focus on the exploration of BERT models that are pretrained and trained from scratch on a site's entire textual data. Although such an approach may demand significant resources and be computationally intensive, it might yield better performance by capturing the unique characteristics and language patterns of a specific health care setting. The exploration of other pretrained language models or more advanced natural language processing techniques could also pave the way for the development of more effective hospitalization prediction methods based on nurse triage notes.

In conclusion, our study demonstrates that while the Bio-Clinical-BERT model does marginally outperform the LR-TF-IDF model in predicting hospitalization from nurse triage notes, the difference is small enough to suggest that simpler methods might be viable for hospitals with limited resources. More research is needed to identify alternative methods that can enhance these models' performance in predicting hospitalization, ultimately improving patient care and hospital resource management.

Through an investigation of the Bio-Clinical-BERT and LR-TF-IDF models' performance, our study contributes to the growing body of literature in the field of NLP and machine learning in health care. It emphasizes the importance of considering the trade-offs between model complexity and performance when deploying predictive tools in clinical settings, highlighting that sometimes, simpler methods can prove as effective as more complex ones.

Conflicts of Interest

None declared.

References

1. Chalfin DB, Trzeciak S, Likourezos A, Baumann BM, Dellinger RP, DELAY-ED study group. Impact of delayed transfer of critically ill patients from the emergency department to the intensive care unit. *Crit Care Med* 2007 Jun;35(6):1477-1483. [doi: [10.1097/01.CCM.0000266585.74905.5A](https://doi.org/10.1097/01.CCM.0000266585.74905.5A)] [Medline: [17440421](https://pubmed.ncbi.nlm.nih.gov/17440421/)]
2. Rabin E, Kocher K, McClelland M, Pines J, Hwang U, Rathlev N, et al. Solutions to emergency department 'boarding' and crowding are underused and may need to be legislated. *Health Aff (Millwood)* 2012 Aug 01;31(8):1757-1766. [doi: [10.1377/hlthaff.2011.0786](https://doi.org/10.1377/hlthaff.2011.0786)] [Medline: [22869654](https://pubmed.ncbi.nlm.nih.gov/22869654/)]
3. Forero R, McCarthy S, Hillman K. Access block and emergency department overcrowding. *Crit Care* 2011 Mar 22;15(2):216 [FREE Full text] [doi: [10.1186/cc9998](https://doi.org/10.1186/cc9998)] [Medline: [21457507](https://pubmed.ncbi.nlm.nih.gov/21457507/)]
4. Sterling NW, Patzer RE, Di M, Schragger JD. Prediction of emergency department patient disposition based on natural language processing of triage notes. *Int J Med Inform* 2019 Sep;129:184-188. [doi: [10.1016/j.ijmedinf.2019.06.008](https://doi.org/10.1016/j.ijmedinf.2019.06.008)] [Medline: [31445253](https://pubmed.ncbi.nlm.nih.gov/31445253/)]
5. Patel D, Cheetirala SN, Raut G, Tamegue J, Kia A, Glicksberg B, et al. Predicting adult hospital admission from emergency department using machine learning: an inclusive gradient boosting model. *J Clin Med* 2022 Nov 22;11(23):634-635 [FREE Full text] [doi: [10.3390/jcm11236888](https://doi.org/10.3390/jcm11236888)] [Medline: [36498463](https://pubmed.ncbi.nlm.nih.gov/36498463/)]
6. Deo RC. Machine learning in medicine. *Circulation* 2015 Nov 17;132(20):1920-1930 [FREE Full text] [doi: [10.1161/CIRCULATIONAHA.115.001593](https://doi.org/10.1161/CIRCULATIONAHA.115.001593)] [Medline: [26572668](https://pubmed.ncbi.nlm.nih.gov/26572668/)]
7. Klang E, Barash Y, Soffer S, Bechler S, Resheff YS, Granot T, et al. Promoting head CT exams in the emergency department triage using a machine learning model. *Neuroradiology* 2020 Feb 10;62(2):153-160. [doi: [10.1007/s00234-019-02293-y](https://doi.org/10.1007/s00234-019-02293-y)] [Medline: [31598737](https://pubmed.ncbi.nlm.nih.gov/31598737/)]
8. Devlin J, Chang MW, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. arXiv Preprint posted online on October 11, 2018.. [doi: [10.48550/arXiv.1810.04805](https://doi.org/10.48550/arXiv.1810.04805)]
9. Soffer S, Glicksberg BS, Zimlichman E, Klang E. BERT for the processing of radiological reports: an attention-based natural language processing algorithm. *Acad Radiol* 2022 Apr;29(4):634-635. [doi: [10.1016/j.acra.2021.03.036](https://doi.org/10.1016/j.acra.2021.03.036)] [Medline: [34362663](https://pubmed.ncbi.nlm.nih.gov/34362663/)]
10. Alsentzer E, Murphy JR, Boag W, Weng WH, Jin D, Naumann T, et al. Publicly available clinical BERT embeddings. arXiv Preprint posted online on April 6, 2019.. [doi: [10.48550/arXiv.1904.03323](https://doi.org/10.48550/arXiv.1904.03323)]
11. Bio_ClinicalBert. Hugging Face. URL: https://huggingface.co/emilyalsentzer/Bio_ClinicalBERT [accessed 2024-08-15]
12. Klang E, Levin MA, Soffer S, Zebrowski A, Glicksberg BS, Carr BG, et al. A simple free-text-like method for extracting semi-structured data from electronic health records: exemplified in prediction of in-hospital mortality. *Big Data Cogn Comput* 2021 Aug 29;5(3):40 [FREE Full text] [doi: [10.3390/bdcc5030040](https://doi.org/10.3390/bdcc5030040)]
13. Khan J, Khondaker MTI, Afroz S, Uddin G, Iqbal A. A benchmark study of machine learning models for online fake news detection. *Mach Learn Appl* 2021 Jun 15;4:100032 [FREE Full text] [doi: [10.1016/j.mlwa.2021.100032](https://doi.org/10.1016/j.mlwa.2021.100032)]
14. Yenikar A, Babu C. AirBERT: a fine-tuned language representation model for airlines tweet sentiment analysis. *Intelligent Decision Technologies* 2023 May 15;17(2):435-455 [FREE Full text] [doi: [10.3233/idt-220173](https://doi.org/10.3233/idt-220173)]
15. Chen S, Li Y, Lu S, Van H, Aerts HJWL, Savova GK, et al. Evaluating the ChatGPT family of models for biomedical reasoning and classification. *J Am Med Inform Assoc* 2024 Apr 03;31(4):940-948. [doi: [10.1093/jamia/ocad256](https://doi.org/10.1093/jamia/ocad256)] [Medline: [38261400](https://pubmed.ncbi.nlm.nih.gov/38261400/)]
16. Xavier BA, Chen P. Natural language processing for imaging protocol assignment: machine learning for multiclass classification of abdominal CT protocols using indication text data. *J Digit Imaging* 2022 Oct;35(5):1120-1130 [FREE Full text] [doi: [10.1007/s10278-022-00633-8](https://doi.org/10.1007/s10278-022-00633-8)] [Medline: [35654878](https://pubmed.ncbi.nlm.nih.gov/35654878/)]
17. Dipnall JF, Lu J, Gabbe BJ, Cosic F, Edwards E, Page R, et al. Comparison of state-of-the-art machine and deep learning algorithms to classify proximal humeral fractures using radiology text. *Eur J Radiol* 2022 Aug;153:110366. [doi: [10.1016/j.ejrad.2022.110366](https://doi.org/10.1016/j.ejrad.2022.110366)] [Medline: [35623313](https://pubmed.ncbi.nlm.nih.gov/35623313/)]
18. Ji S, Hölttä M, Martinen P. Does the magic of BERT apply to medical code assignment? a quantitative study. *Comput Biol Med* 2021 Dec;139:104998 [FREE Full text] [doi: [10.1016/j.combiomed.2021.104998](https://doi.org/10.1016/j.combiomed.2021.104998)] [Medline: [34739971](https://pubmed.ncbi.nlm.nih.gov/34739971/)]
19. Zhang X, Kim J, Patzer RE, Pitts SR, Patzer A, Schragger JD. Prediction of emergency department hospital admission based on natural language processing and neural networks. *Methods Inf Med* 2017 Oct 26;56(5):377-389. [doi: [10.3414/ME17-01-0024](https://doi.org/10.3414/ME17-01-0024)] [Medline: [28816338](https://pubmed.ncbi.nlm.nih.gov/28816338/)]
20. Raita Y, Goto T, Faridi MK, Brown DFM, Camargo CA, Hasegawa K. Emergency department triage prediction of clinical outcomes using machine learning models. *Crit Care* 2019 Feb 22;23(1):64 [FREE Full text] [doi: [10.1186/s13054-019-2351-7](https://doi.org/10.1186/s13054-019-2351-7)] [Medline: [30795786](https://pubmed.ncbi.nlm.nih.gov/30795786/)]
21. Klang E, Kummer BR, Dangayach NS, Zhong A, Kia MA, Timsina P, et al. Predicting adult neuroscience intensive care unit admission from emergency department triage using a retrospective, tabular-free text machine learning approach. *Sci Rep* 2021 Jan 14;11(1):1381 [FREE Full text] [doi: [10.1038/s41598-021-80985-3](https://doi.org/10.1038/s41598-021-80985-3)] [Medline: [33446890](https://pubmed.ncbi.nlm.nih.gov/33446890/)]

Abbreviations

AUC: area under the receiver operating characteristic curve
BERT: Bidirectional Encoder Representations from Transformers
Bi-LSTM: bidirectional long short-term memory
BOW: bag-of-words
ED: emergency department
LR: logistic regression
MI: mutual information
MIMIC: Medical Information Mart for Intensive Care
MSHS: Mount Sinai Health System
MSQ: Mount Sinai Queens
NLP: natural language processing
OR: odds ratio
ROC: receiver operating characteristic
TF-IDF: term frequency–inverse document frequency
W2V: Word-2-Vec
XGBoost: extreme gradient boosting

Edited by K El Emam, B Malin; submitted 25.08.23; peer-reviewed by M Elbattah, LH Yao; comments to author 08.02.24; revised version received 15.04.24; accepted 14.06.24; published 27.08.24.

Please cite as:

Patel D, Timsina P, Gorenstein L, Glicksberg BS, Raut G, Cheetirala SN, Santana F, Tamegue J, Kia A, Zimlichman E, Levin MA, Freeman R, Klang E

Traditional Machine Learning, Deep Learning, and BERT (Large Language Model) Approaches for Predicting Hospitalizations From Nurse Triage Notes: Comparative Evaluation of Resource Management
JMIR AI 2024;3:e52190

URL: <https://ai.jmir.org/2024/1/e52190>

doi: [10.2196/52190](https://doi.org/10.2196/52190)

PMID:

©Dhavalkumar Patel, Prem Timsina, Larisa Gorenstein, Benjamin S Glicksberg, Ganesh Raut, Satya Narayan Cheetirala, Fabio Santana, Jules Tamegue, Arash Kia, Eyal Zimlichman, Matthew A Levin, Robert Freeman, Eyal Klang. Originally published in JMIR AI (<https://ai.jmir.org>), 27.08.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR AI, is properly cited. The complete bibliographic information, a link to the original publication on <https://www.ai.jmir.org/>, as well as this copyright and license information must be included.

Original Paper

Obtaining the Most Accurate, Explainable Model for Predicting Chronic Obstructive Pulmonary Disease: Triangulation of Multiple Linear Regression and Machine Learning Methods

Arnold Kamis^{1*}, PhD; Nidhi Gadia^{1*}, MS; Zilin Luo^{1*}, MS; Shu Xin Ng^{1*}, MS; Mansi Thumbar^{1*}, MS

Brandeis International Business School, Brandeis University, Waltham, MA, United States

* all authors contributed equally

Corresponding Author:

Arnold Kamis, PhD

Brandeis International Business School

Brandeis University

Sachar International Center

415 South St

Waltham, MA, 02453

United States

Phone: 1 781 736 8544

Fax: 1 781 736 2269

Email: akamis@brandeis.edu

Abstract

Background: Lung disease is a severe problem in the United States. Despite the decreasing rates of cigarette smoking, chronic obstructive pulmonary disease (COPD) continues to be a health burden in the United States. In this paper, we focus on COPD in the United States from 2016 to 2019.

Objective: We gathered a diverse set of non-personally identifiable information from public data sources to better understand and predict COPD rates at the core-based statistical area (CBSA) level in the United States. Our objective was to compare linear models with machine learning models to obtain the most accurate and interpretable model of COPD.

Methods: We integrated non-personally identifiable information from multiple Centers for Disease Control and Prevention sources and used them to analyze COPD with different types of methods. We included cigarette smoking, a well-known contributing factor, and race/ethnicity because health disparities among different races and ethnicities in the United States are also well known. The models also included the air quality index, education, employment, and economic variables. We fitted models with both multiple linear regression and machine learning methods.

Results: The most accurate multiple linear regression model has variance explained of 81.1%, mean absolute error of 0.591, and symmetric mean absolute percentage error of 9.666. The most accurate machine learning model has variance explained of 85.7%, mean absolute error of 0.456, and symmetric mean absolute percentage error of 6.956. Overall, cigarette smoking and household income are the strongest predictor variables. Moderately strong predictors include education level and unemployment level, as well as American Indian or Alaska Native, Black, and Hispanic population percentages, all measured at the CBSA level.

Conclusions: This research highlights the importance of using diverse data sources as well as multiple methods to understand and predict COPD. The most accurate model was a gradient boosted tree, which captured nonlinearities in a model whose accuracy is superior to the best multiple linear regression. Our interpretable models suggest ways that individual predictor variables can be used in tailored interventions aimed at decreasing COPD rates in specific demographic and ethnographic communities. Gaps in understanding the health impacts of poor air quality, particularly in relation to climate change, suggest a need for further research to design interventions and improve public health.

(JMIR AI 2024;3:e58455) doi:[10.2196/58455](https://doi.org/10.2196/58455)

KEYWORDS

chronic obstructive pulmonary disease; COPD; cigarette smoking; ethnic and racial differences; machine learning; multiple linear regression; household income; practical model

Introduction

Background

Lung disease is a severe problem in the United States. According to the Centers for Disease Control and Prevention (CDC), asthma is responsible for at least 3000 deaths per year and chronic obstructive pulmonary disease (COPD) is responsible for at least 150,000 deaths per year. COPD is a progressive lung disease, encompassing chronic bronchitis and emphysema, which is characterized by airflow limitation and breathing difficulties. Asthma and COPD can co-occur (asthma-COPD overlap), with increased risk of mortality [1] and diminished disease-related quality of life [2]. This is from a variety of factors, some under individual control, such as cigarette smoking, and others not under individual control, such as ambient air pollution.

Cigarette smoking has been trending downward in recent years, thanks in part to public health advertisement campaigns. Nevertheless, air quality can be dangerously poor at times, which exacerbates lung health problems [3], and the impacts can be particularly acute in populations considered vulnerable. Technologically, there are tools that help individuals avoid poor air quality. For example, there are mobile phone apps that track air quality. They notify their owners on days when air quality is dangerously poor, advising them to stay indoors or avoid strenuous outdoor exercise. The effectiveness of such apps is ambiguous thus far [4,5].

The rest of the paper is organized as follows. We first review prior work regarding the possible factors contributing to COPD in adults. We then describe our methods, including data sources for the variables of interest and descriptive statistics. Following this, we will describe and interpret the results of our multiple linear regression (MLR) and machine learning (ML) models. We conclude by describing the overall research contributions as well as limitations and future directions.

Prior Work

There is substantial literature on factors contributing to COPD, including a wide variety of environmental, economic, and demographic variables; the etiology of COPD is multifactorial, with smoking being the most well-known contributing factor. Furthermore, the combination of environmental pollutants and cigarette smoke has shown synergistic effects, accelerating the decline in lung function and worsening COPD [6,7]. In addition, occupational exposures, for example, to coal dust, arsenic, or diesel fumes, or to home exposures, such as gas stoves, wood stoves, kerosene heaters, and fireplaces, contribute to overall COPD outcomes. When combined with persistent ambient air pollution, the risk and severity of COPD will likely increase [8].

Pollutants and copollutants are associated with decreased lung function and can lead to COPD. The loss can range from mild, such as allergies, to severe, that is, mortality. Air quality varies widely throughout the United States because of pollutants and copollutants, and climate change may be worsening it, particularly for populations considered vulnerable [9]. Health disparities due to poor quality air and other stressors are well

known [10-12]. Ambient air pollution in poorer neighborhoods tend to be exacerbated by additional copollutants, heat stress, and aeroallergens. Air quality index (AQI) includes the totality of pollutants and copollutants.

ML methods have been applied increasingly to public health and medical problems. For example, ML has been used to support the public health response to COVID-19 through surveillance, case identification, contact tracing, and evaluating interventions [13]. ML methods have been used as a supportive tool to recognize cardiac arrest in emergency calls [14]. In that study, Zicari et al [14] developed a general protocol with a collaborative team to ensure that the ML tool was domain- and context-sensitive as well as abiding by ethical guidelines, thus obtaining trustworthiness. ML has been also used to improve early and accurate stroke recognition during emergency medical calls [15].

ML methods have been used to study COPD, in particular. For example, ML methods have been used to develop a prediction system using lifestyle data, environmental factors, and patient symptoms for the early detection of acute exacerbations of COPD within a 7-day window [16]. Another study on acute exacerbations of COPD compared several ML methods and found that a decision tree classifier was best for assessing patient severity and guiding treatment strategy [17]. In another study, to improve mortality prediction from COPD, a random forest was used to identify the most important imaging features [18]. Gradient boosted trees (GBTs) have been used to predict lung function values from computed tomography images obtained from patients with COPD and those without COPD [19]. Deep learning has been effective in analyzing images diagnostic of COPD [20]. Finally, research using a generalized linear model found a complex relationship between rural living and COPD-related outcomes in US veterans [21]. Thus, a variety of ML models have been successfully applied for use in public health scenarios in general and COPD in particular. The one that ultimately works best in a given situation depends on many factors.

Different races and ethnicities may have different baseline rates of disease due to various factors, including historical misdiagnosis and mistreatment of various racial or ethnic groups, which leads to differential outcomes [22]. There may be outcome, equity, and counseling differences by gender as well as race or ethnicity in the diagnosis and treatment of COPD [23,24].

We had three general expectations of COPD in our models:

1. Cigarette smoking will have the highest impact on COPD rates.
2. AQI will have a strong impact on COPD rates.
3. There will be differences in COPD rates based on racial or ethnic demographics.

Methods

Overview

This paper used MLR and ML methods to predict COPD at the core-based statistical area (CBSA) level [25]. At the time of

this study, there were 388 metropolitan and 541 micropolitan statistical areas in the United States. The data sources were obtained from data repositories of 3 official US agencies, specifically from the CDC. We gathered, integrated, and checked them for data quality. By combining different variables from this variety of data sources, we aimed to obtain a uniquely high

accuracy model, while simultaneously reducing biases or flaws that may be attributable to individual data sources. We further checked for missing values (ie, NULL or NA) in every variable. We checked for data correctness by checking the plots of the distributions for every variable, looking for impossible or outlying values. [Table 1](#) shows the data sources used.

Table 1. Data sources.

Source	Reference
National Center for Health Statistics	[26]
Chronic Disease Indicators data	[27]
US Chronic Disease Indicator, stratification values	[28]

Data were collected for all CBSAs that were available from 2016 to 2019. All data obtained from the CDC were contributed voluntarily at the individual level and aggregated to remove all personally identifiable information [29].

The COPD rates are for 2019, whereas all the predictor variables are averaged over the timespan from 2016 to 2018. As such, the models obtained are predictive over time. The data collection result was 517 (56%) of the 929 CBSAs, with proportionally more from the 388 metropolitan statistical areas than from the 541 micropolitan statistical areas. The response variable is the

percentage of the CBSA having COPD. We modeled all factors as random variables directly contributing to COPD, which is measured as the proportion (percentage) of the population having COPD. Race or ethnicity was also modeled as percentage of the population rather than as categorical variables. All variables in [Table 2](#) are averaged as mean, except for household income, which was averaged as median.

In [Figure 1](#), we observe that some variables (ie, population, gross domestic product [GDP], GDP per capita, and median household income) are skewed in their distribution.

Table 2. Main variables and descriptive statistics and average within core-based statistical areas.

	Years	Values, median (IQR)	Values, mean (SD)	Values, range
Population (n)	2016-2018	96,811 (48,763-180,484)	191,892 (408,308)	7351-6,633,096
GDP ^a (US \$)	2016-2018	13,126,907 (2,562,704-39,046,120)	64,223,036 (212,975,821)	447,355-3,218,209,695
Median household income (US \$)	2016-2018	52,632 (46,867-60,494)	54,736 (11,319)	27,842-119,332
GDP per capita (US \$)	2016-2018	100.07 (47.83-277.17)	253.77 (479)	16.86-4731.50
Air quality index	2016-2018	38.67 (34.00-43.00)	38.02 (10)	9.00-95.00
Smoking rate	2016-2018	17.12 (15.33-19.28)	17.29 (3)	8.41-29.59
Poverty rate (all ages)	2016-2018	13.80 (10.92-17.12)	14.36 (4)	3.87-35.56
Unemployment rate	2016-2018	4.52 (3.67-5.43)	4.71 (2)	1.97-20.93
Education rate	2016-2018	22.91 (17.94-27.96)	24.22 (8)	8.77-65.75
White (%)	2016-2018	87.6 (78.6-92.8)	84.6 (0.129)	22.1-100
Black (%)	2016-2018	4 (1.5-12.5)	9.3 (0.124)	0.3-100
AI or AN ^c (%)	2016-2018	0.7 (0.4-1.7)	2 (0.044)	0.1-45.9
Asian (%)	2016-2018	1.6 (0.9-3)	2.8 (0.041)	0.2-42.8
NH or PI ^d (%)	2016-2018	0.1 (0.1-0.2)	0.3 (0.01)	0.0-12.9
Hispanic (%)	2016-2018	7 (3.9-14.9)	13.3 (0.164)	0.9-95.5
COPD ^b rate (%)	2019	6.7 (5.7-7.9)	6.871 (1.511)	3.2-15

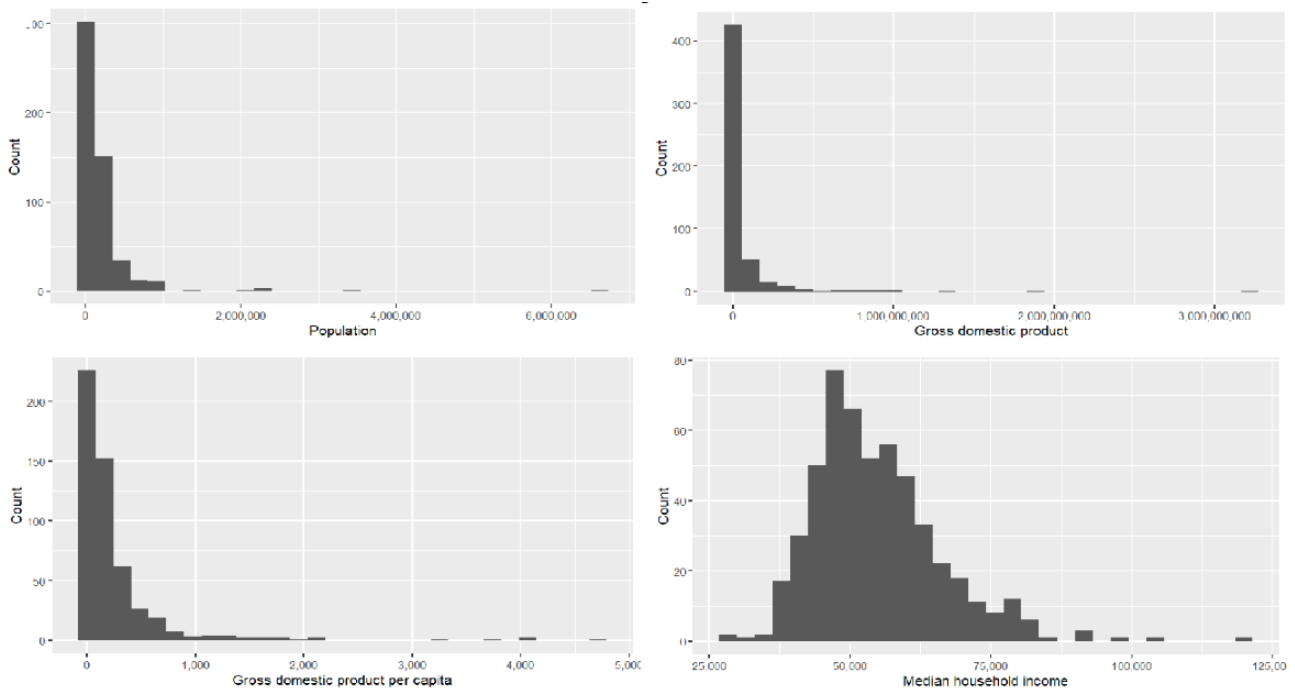
^aGDP: gross domestic product.

^bCOPD: chronic obstructive pulmonary disease.

^cAI or AN: American Indian or Alaska Native.

^dNH or PI: Native Hawaiian or Pacific Islander.

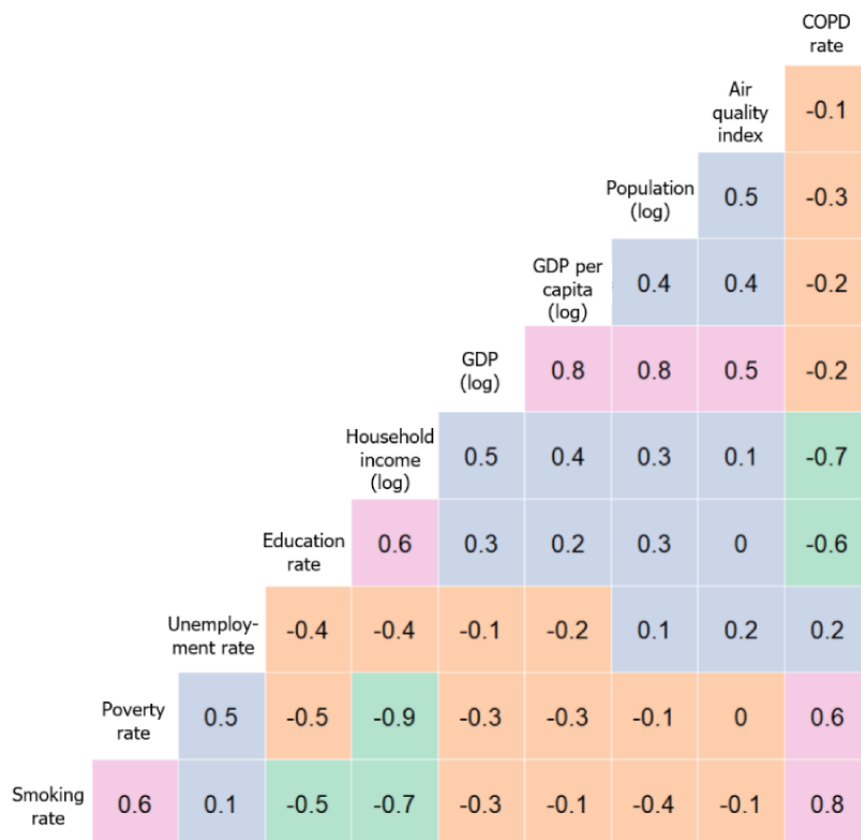
Figure 1. Population, gross domestic product (GDP), GDP per capita, and median household income.



Therefore, we made a log transformation of these variables (ie, logPopl, logGDP, logGDPpc, and logHHI) to make them less

skewed, and we show a heat map of correlations of them with the other variables in Figure 2.

Figure 2. Correlations among main variables. COPD: chronic obstructive pulmonary disease; GDP: gross domestic product.



We see a range of correlations, from very negative (green) to negative (orange) to positive (purple) to very positive (pink). In the rightmost column, we see the correlations between the response variable, COPD rate, and the other variables, ranging

from very positive (smoking rate) to moderately positive (poverty and unemployment rates) to moderately negative (education and logged household income) to slightly negative (log of GDP, log of population, log of GDP per capita, and

AQI). Given these correlations, we are likely to find good predictive models, but we need to check for multicollinearity in any linear model that we identify.

To understand and model COPD, one has to consider the consistently largest contributing factor: cigarette smoking. Research tends to either control for cigarette smoking or exclude it entirely. In this paper, we chose to include cigarette smoking, accounting for it in our models, but also to examine other factors to compare the magnitudes of influence among the various

factors. We aimed to model a variety of factors, including cigarette smoking, to arrive at the model that predicts COPD with the greatest accuracy.

Statistical Analysis

Overview

Our MLR baseline model in R (version 4.2.3) yielded the output in Table 3, which is sorted by absolute value of the *t* value, from high to low.

Table 3. Multiple linear regression.

	Estimate	SE	<i>t</i> test (<i>df</i> =503)	<i>P</i> value
(Intercept)	32.4000	2.930	11.065	<.001
Smoking_Rate	0.2570	0.015	16.635	<.001
Log_HH_Income	-2.8100	0.264	-10.638	<.001
Hispanic_percentage	-2.3900	0.234	-10.249	<.001
Education_Rate	-0.0334	0.005	-6.627	<.001
AI_or_AN_percentage ^a	-2.9000	0.778	-3.726	<.001
Black_percentage	-1.2700	0.356	-3.558	<.001
NH_or_PI_percentage ^b	-15.8000	4.430	-3.558	<.001
Log_GDP	0.0741	0.035	2.126	.034
White_percentage	-0.7380	0.358	-2.060	.04
Unemployment_Rate	0.0456	0.023	1.993	.047
Asian_percentage	2.4800	1.250	1.988	.047
Log_Population	0.0899	0.055	1.626	.105
Air_Quality_Index	0.0003	0.003	0.098	.922

^aAI_or_AN: American Indian or Alaska Native.

^bNH_or_PI: Native Hawaiian or Pacific Islander.

The model has residual SE 0.658 on 503 *df*. The multiple R^2 is 0.8152 and adjusted R^2 is 0.8105. The *F*-statistic is 170.7 on 13 and 503 *df* ($P<.001$). The variance inflation factors were checked, with all values <5 indicating low multicollinearity.

There are 7 predictors of high statistical significance: smoking rate, Black percentage, Native Hawaiian or Pacific Islander percentage, American Indian or Alaska Native percentage, education rate, Hispanic percentage, and log of household income. Smoking rate has a positive association with COPD, with every additional percentage increase associated with a 0.257% increase in the COPD rate. The other 6 highly significant predictors have a negative association. Every percentage increase in the log of household income lowers the COPD rate by 2.81%. The Hispanic percentage is nearly as strong; every percentage increase corresponds to a drop of 2.39% in COPD rate. American Indian or Alaska Native is a bit stronger in its coefficient estimate; every percentage point increase corresponds to a drop of 2.9% in COPD rate. Every percentage point increase in Native Hawaiian or Pacific Islanders corresponds to a drop of 15.8%, which is much stronger. Every percentage point increase in Black percentage corresponds to a drop of 1.27% in COPD rate. Education rate has a strongly statistically significant relationship, but a small

percentage point impact: every percentage increase corresponds to a decrease of 0.0334% in COPD rate. The remaining 4 predictors—White percentage, GDP (logged), unemployment rate, and Asian percentage—are far less statistically significant and, therefore, should be interpreted with caution.

Linear models are simpler than ML models, and they are sometimes perfectly adequate for explaining a phenomenon. They are easier to interpret, communicate, and implement as new policy. They make statistical assumptions, which can be verified. Linear regression is certainly a good place to start. However, we argue that one should not stop there because an ML model can capture substantial variance from nonlinear relationships (if there are any) in the data and thus produce a more accurate model. By capturing additional variance, the model can capture subtler effects and relationships due to interactions, context, and tipping points. This is crucial because public health practice tends to use simple if-then rules, that is, decision trees. ML models can add nuance to those decision trees based on the captured nonlinearities. Although an adjusted R^2 of 0.8105 looks quite strong, we can perhaps do better with ML methods [18-21].

The 7 ML methods evaluated in this paper are lasso regression, ridge regression, generalized additive model, support vector

machine, artificial neural network, random forest, and GBT. These methods were selected for their known strengths in minimizing errors of bias or errors of variance, that is, their ability to fit data well on test data without overfitting. They also represent the range of algorithms commonly used in ML prediction, from methods established in classical statistics to more modern methods derived from computer science. They are commonly used because they are accurate and well understood. Trying a variety of methods is a common practice because the different methods make different statistical assumptions, which may enhance or inhibit optimal performance. All methods were available as R packages for R (version 4.2.3). We summarize each method in terms of its main pros and cons:

Lasso Regression (L1 Regularization)

Lasso regression is an MLR method that incorporates regularization to perform variable selection. It minimizes the sum of squared errors between predicted and actual values, while adding a penalty term based on the absolute value of coefficients multiplied by a tuning parameter. Doing so shrinks some coefficients to exactly 0, effectively performing feature selection by excluding less important variables from the model. This reduces model complexity and minimizes multicollinearity. This is a standard refinement of MLR (R package glmnet).

Ridge Regression (L2 Regularization)

Ridge regression is an MLR technique that adds a penalty term to the objective function to reduce the coefficients of less important predictors and guard against overweighting the most important predictors. While it retains all predictors in the model, ridge regression can help improve the robustness of the model in the presence of correlated predictors by reducing multicollinearity. This is a standard refinement of MLR (R package ridge).

Generalized Additive Model

The generalized additive model is a nonparametric generalization of MLR, which allows for nonlinear terms and coefficient regularization while maintaining interpretability. Each term is a function of X_n rather than simply a numeric coefficient multiplied with X_n . As with MLR, all the terms are added together. Although overfitting can occur, regularization and cross-validation help to minimize it (R package mgcv).

Support Vector Machine

Support vector machine is a technique that transforms the data into a high-dimensional variable space using a kernel function, fitting a function that best fits the data while allowing a certain margin of error (epsilon) and maintaining robustness against outliers. Epsilon tubes can provide a visual representation of the model's uncertainty. Points within the tube are considered well predicted, while those outside represent errors. A regularization parameter controls the trade-off between accuracy and complexity (R package e1071).

Artificial Neural Network

Artificial neural network is a generalization of MLR with hidden layers of nodes between input and output nodes; it may result in overfitting. Depending on the number of hidden layers, nodes

per layer, and the activation function used to convert inputs to outputs, an arbitrarily complex model can be fit. This can be thought of as a simplified version of a human brain, in which input and output nodes are separated by ≥ 1 layers of hidden nodes. Prediction error causes the weights of the hidden nodes to be adjusted until minimal error is achieved (R package neuralnet).

Random Forest

Random forest is an ensemble technique to fit a large number of a bootstrap-sampled aggregation (bagging) of trees by considering a random subset of variables at each tree split. Intuitively, a random forest is a blending of a large number of decision trees, the "wisdom of the forest." The random subset of variables restriction is done to prevent strong variables from dominating the weaker variables. A random forest tends to perform very well but is difficult to interpret (R package RandomForest).

Gradient Boosted Trees

GBT is an ensemble of sequential trees that focuses on the errors of the previous tree. It is able to find interaction effects implicitly. It uses gradient descent search to rapidly minimize error via an arbitrary, differentiable loss function. It uses many trees to help ensure that the local minimum error found is the global minimum. Intuitively, this builds a strong predictive model by combining many weak models, each correcting the errors of the previous one (R package XGBoost).

Our ML approach followed best practices. We randomly partitioned the data set into train (311/517, 60%), cross-validate (103/517, 20%), and test (103/517, 20%) subsets. We checked for outliers, multicollinearity, and target leakage to ensure valid models [30].

Ethical Considerations

This research did not involve human subjects at the individual level and therefore did not require institutional review board approval. Our data were collected from CDC sources at the level of CBSA. All sources were free of personal identifying information, because the CDC is legally required to ensure the protection of the data. All data were collected and aggregated in a non-personal identifying information manner. The results of our analysis do suggest communicating with different racial and ethnic groups differently, tailoring the implications directly to patients as well as indirectly to their families, communities, and health care providers in a race- or ethnicity-sensitive manner.

Results

In Table 4, we describe the results of the ML models of COPD by various accuracy metrics. For the accuracy metrics, we used 3 standard measures of predictive accuracy in addition to variance explained (adjusted R^2): root mean square error (RMSE), mean absolute error (MAE), and symmetric mean absolute percentage error (SMAPE) [31,32]. We performed a grid search over all the main numeric parameters for a given method to find the optimal combination of parameter values [33]. A grid search tries all combinations of parameters from a

minimum to a maximum value by some step size. Those minimum, maximum, and step sizes are determined from typical

default values and best practices. The best metrics in Table 4 are indicated by italics.

Table 4. Machine learning models versus multiple linear regression.

Method	Adjusted R^2	Root mean square error			Mean absolute error			Symmetric mean absolute percentage error		
		Train	CV ^a	Test	Train	CV	Test	Train	CV	Test
Gradient boosted tree (XGBoost, loss function=least squares, learning rate=0.05, and maximum tree depth=10)	0.857	0.550	0.598	0.557	0.433	0.445	<i>0.456</i> ^b	6.473	6.543	<i>6.956</i>
Support vector machine (Nystroem kernel and loss function=Poisson deviance)	<i>0.858</i>	0.555	<i>0.558</i>	<i>0.556</i>	0.435	<i>0.434</i>	0.462	6.515	<i>6.443</i>	6.989
Random forest (maximum trees=500, maximum depth=none, and maximum leaves=100)	0.836	<i>0.534</i>	0.614	0.596	<i>0.420</i>	0.462	0.479	<i>6.315</i>	6.819	7.339
Neural network (2 layers: 512, 512 units; regularization via random dropout rate=0.05 and activation function=prelu)	0.845	0.601	0.609	0.580	0.455	0.467	0.468	6.856	6.928	7.182
Generalized additive model (learning rate=0.3, maximum bins=100, and loss function=least squares)	0.822	0.629	0.658	0.621	0.515	0.508	0.488	7.619	7.502	7.212
Ridge regression	0.810	0.589	0.618	0.641	0.467	0.483	0.527	6.986	7.346	7.986
Lasso regression	0.758	0.750	0.778	0.724	0.585	0.593	0.597	8.425	8.544	8.824
Multiple linear regression	0.811	0.620	0.699	0.749	0.474	0.548	0.591	7.205	8.403	9.666

^aCV: cross-validation.

^bValues in italics represent the best metrics.

The ML methods were superior to MLR on most metrics. Support vector machine was the best on adjusted R^2 and RMSE, slightly superior to GBT, but GBT was superior by a larger margin on MAE and SMAPE. Therefore, we chose GBT as the best overall method. In [Multimedia Appendix 1](#), we show the variable importance plot for the GBT model. Variable importance plots are a common first way to peer inside a “black-box method” and understand the relative importance of the variables used within it [34].

The top five variables in terms of impact were (1) smoking rate and (2) household income, followed by (3) American Indian or Alaska Native percentage, (4) education rate, and (5) unemployment rate. Black percentage was sixth, Hispanic percentage was seventh, and there was only a small impact from the remaining variables: White percentage, AQI, Asian

percentage, Native Hawaiian or Pacific Islander percentage, population, and GDP. Relative to the MLR, smoking rate, household income, education rate, and Black percentage remained the same in terms of rank importance. Hispanic percentage dropped from third to seventh rank; American Indian or Alaska Native percentage rose from fifth to third rank; and unemployment rate rose sharply, from 10th to 5th in importance. Native Hawaiian or Pacific Islander percentage dropped sharply, from 7th to 11th in rank.

[Figure 3](#) shows the lift plot, and [Figure 4](#) shows the predictive residual plot. The lift plot shows observations sorted by predicted value deciles. The ratio of the observed outcome to the expected outcome was calculated and plotted. The predictive residual plot shows the differences between observed and predicted values.

Figure 3. Lift plot showing chronic obstructive pulmonary disease rate as a function of 10 decile bins; predicted values are in blue and actual values are in red. COPD: chronic obstructive pulmonary disease.

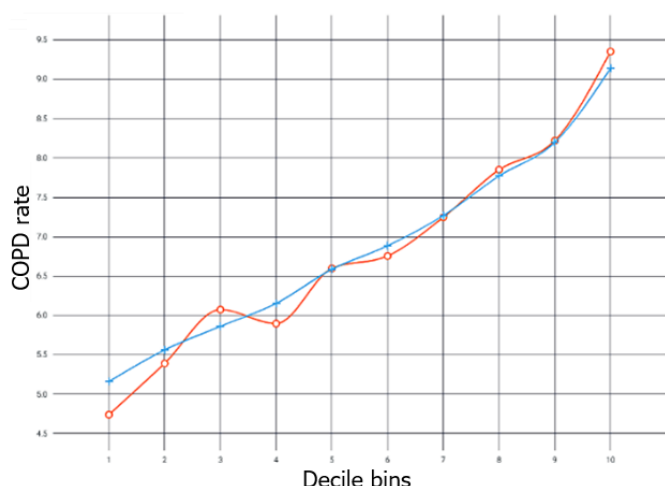
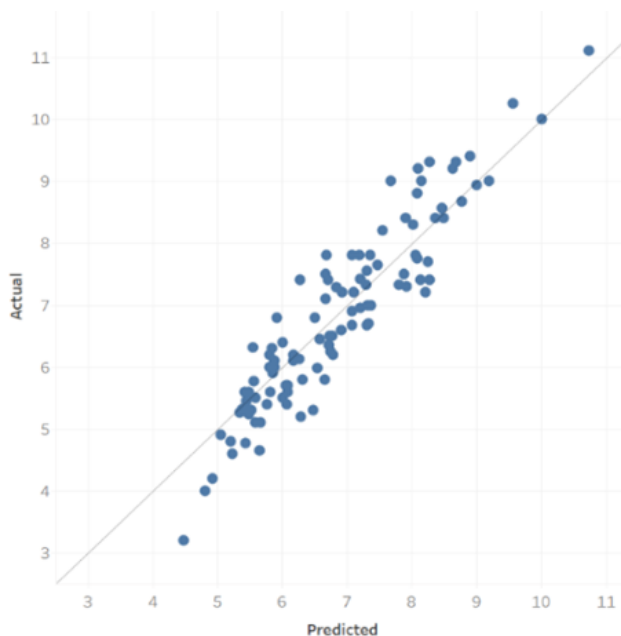


Figure 4. Prediction residuals.



In addition to the variable importance plot, other plots were used to gain an understanding of ML models: local interpretable model-agnostic explanations (LIME) models and SHAP (Shapley additive explanations) plots [35-37]. We chose SHAP plots because they are based on a cooperative game-theoretical foundation, showing every combination of the variables in the model and how they work together to predict the outcome variable. Figure 5 shows the SHAP plot for all the GBT’s variables.

The top 5 variables (smoking rate, household income, American Indian or Alaska Native percentage, education rate, and unemployment rate) have substantially more impact on COPD percentage than the remaining variables. We show the top 5 variables as well as the next 4 as individual SHAP plots of the GBT in Figure 6. All 9 plots show significant nonlinearities.

Smoking had the greatest impact: as the smoking rate increased, the COPD rate also rose substantially, following a steeply curved, nearly exponential relationship. Median household income had the second highest impact, an almost linear (and negative) relationship. The greater the household income, the lower the COPD rate. This could indicate better insurance coverage, better health care access, higher quality health care (ie, prevention or treatment), lower occupational exposure, or lower home exposure (eg, gas stoves). The next variable was American Indian or Alaska Native percentage, indicating a negative but nonlinear relationship with COPD rate: a steep drop followed by a gradual tapering. This represents a significant protective influence shown for the American Indian or Alaska Native community, which has not yet been noted in the literature.

Figure 5. SHAP (Shapley additive explanations) values for all features (variables). AI: American Indian; AN: Alaska Native; GDP: gross domestic product; NH: Native Hawaiian; PI: Pacific Islander.

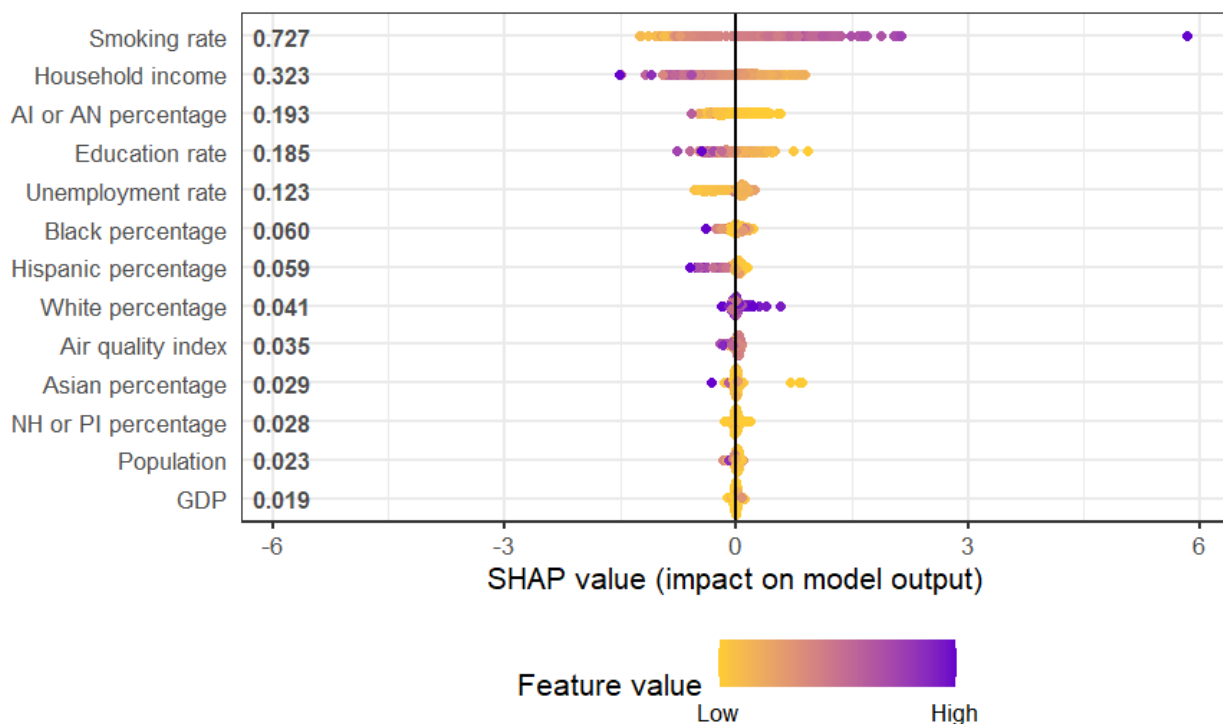
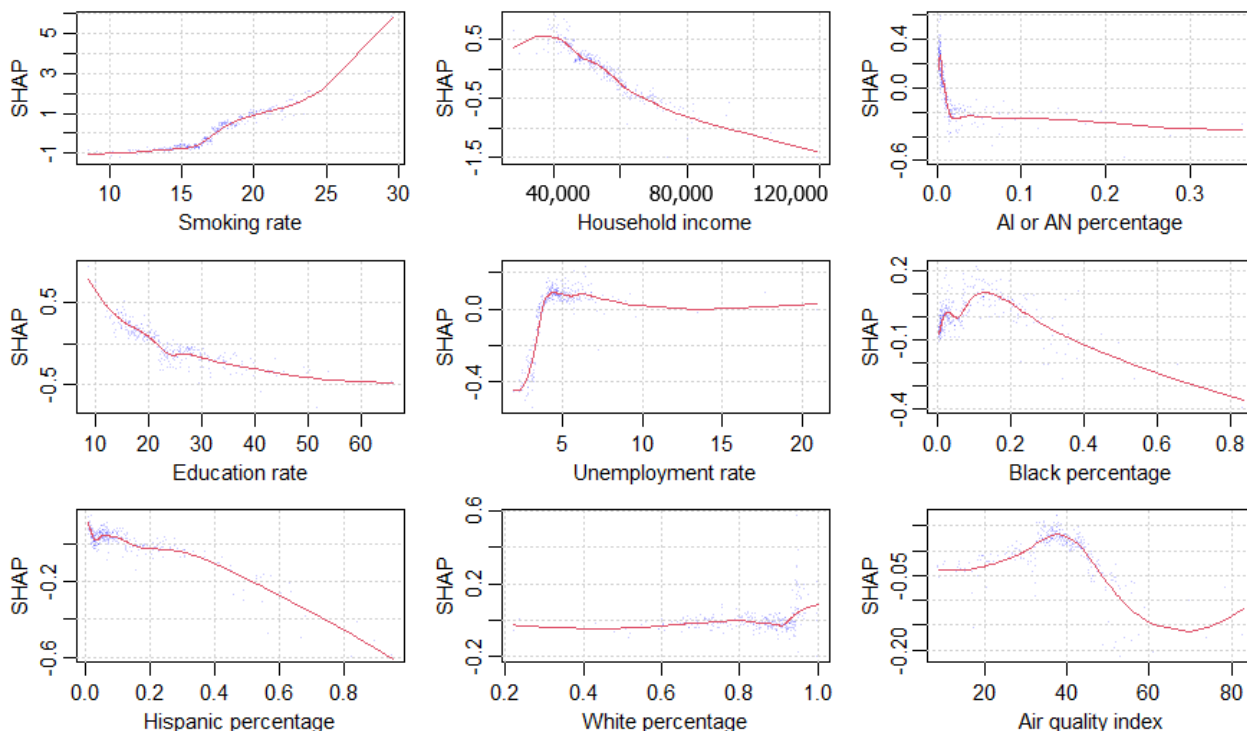


Figure 6. SHAP (Shapley additive explanations) plots for the 9 most important variables. AI: American Indian; AN: Alaska Native.



The next variable, education rate had a negative, curvilinear relationship. The more educated the population, the lower the COPD rate. The explanation could be similar to that of income: better insurance coverage or health care access, better quality of health care, lower occupational exposure, or lower home exposure [38]. The next variable was unemployment rate, with a sharply positive but flat relationship with COPD rate. The

next variable was Black percentage, with an initial positive relationship with COPD rate but then a reversal to a negative, linear relationship.

The next variable, Hispanic percentage, showed a negative linear relationship with COPD rate. This represents a significant protective influence shown for people in the Hispanic

community, which is consistent with the literature [39-45]. The next variable was White percentage, showing a slightly negative relationship with COPD rate. Finally, the last variable was AQI (higher value being worse), which shows an initial positive relationship with COPD rate, peaking around 38. This may be a critical point, after which people take precautions not to be exposed to the low-quality air.

Discussion

Principal Findings

We had three general expectations, which were largely met:

1. The impact of cigarette smoking was the largest in all models.
2. The AQI had an impact in the best ML model, but it was smaller than expected.
3. There were substantial racial or ethnic differences, particularly among American Indian or Alaska Native, Black, and Hispanic communities.

Consistent with the literature, we found that smoking remains the most significant risk factor for COPD, with research consistently demonstrating a strong association between smoking status and COPD prevalence. In our MLR, we found that smoking rate is the strongest predictor of COPD rate. We found the same result in our GBT but also found that the smoking rate has a curvilinear, almost exponential, relationship with COPD. The Rotterdam study, a large-scale population-based cohort study, found that current and former smokers had a substantially higher risk of developing COPD compared to never smokers [46]. A nationwide population-based cohort study in South Korea demonstrated that smoking cessation after COPD diagnosis was associated with lower all-cause and cause-specific mortality [47].

Notably, 3 of the 4 next most important variables, in terms of impact in our GBT, are socioeconomic variables: household income (rank 2), education rate (rank 4), and unemployment rate (rank 5). In the MLR, we found that household income (logged) had the second highest impact. In the GBT, household income had the second highest impact, but the tipping point was around US \$40,000, after which higher income had a linear, negative relationship with COPD. Education rate had a strongly negative, curvilinear relationship with COPD. Unemployment rate had a sharply positive relationship with COPD, but then peaked at 5% unemployment, after which it plateaued.

These results are largely consistent with the literature on socioeconomic factors and smoking behavior, suggesting an indirect relationship with COPD via smoking. A study examining smoking among adolescents in 6 European cities found that disposable income was positively associated with smoking [48]. Conversely, lower socioeconomic status was associated with higher COPD prevalence because in addition to lower education and income, there may be environmental pollutants, occupational hazards, or barriers to COPD screening, diagnosis, and treatment [49]. In contrast with the literature, our SHAP plots show mostly nonlinear relationships with COPD. Household income showed a tipping point at US

\$40,000, after which the negative relationship with COPD was nearly linear.

Ethnic or racial variables accounted for 3 of the top 7 variables in the GBT: American Indian or Alaska Native percentage (rank 3), Black percentage (rank 6), and Hispanic percentage (rank 7). The greater the size of those minority populations, the lower the COPD rate. Our SHAP plots show significant tipping points (nonlinearities) for American Indian or Alaska Native percentage and Black percentage and a mostly linear relationship for Hispanic percentage. Consistent with the literature, all 3 variables show a strongly negative association with COPD.

The regression and GBT models show that in addition to strongly protective impacts for lower cigarette smoking and higher household income, there are protective impacts for larger American Indian or Alaska Native and Hispanic populations as well as a nonlinear impact on larger Black populations. Higher education rate and lower unemployment rate are also protective, whereas AQI shows mixed effects. These results have implications for private health care practitioners, public health care officials, and health care policy makers who aim to reduce COPD rates. Such policies and programs should not assume high digital literacy [50,51]. System designers could use SMS text messaging, social media, and interactive voice response systems. This would be appropriate for those with lower household income or lower education levels. To design culturally appropriate visual cues and messaging to different racial or ethnic groups, members of the various communities should be included in the design process [52,53]. In sum, the user interface should exhibit high ease-of-use—using gamification, storytelling, and peer support—consistent with cultural norms.

Several studies have identified ethnic and racial disparities in COPD prevalence and risk among smokers. One study found that racial and ethnic minority individuals, particularly African Americans and Hispanics, had a lower prevalence of airflow obstruction than non-Hispanic White individuals, even after adjusting for smoking status and other risk factors [54]. This finding was supported by another study that observed lower COPD risk in ethnic minority groups compared to White individuals, despite similar smoking intensities [55]. A larger minority population means a larger peer support network for prevention and cessation of smoking and a larger peer community to recommend COPD screening, diagnosis, and treatment, which is particularly useful in a health care system that has implicit racial or ethnic bias [50,56].

There are varying levels of patient trust and implicit bias in health care practitioners themselves [57], which contributes to health outcome differences. From a population communication perspective, messaging regarding the risks of COPD—particularly the avoidance or cessation of cigarette smoking—should be sensitive to community context, engaging trusted local authorities to optimize the chances of patient engagement [58]. Health care practitioners could partner with trusted local authorities and community leaders regarding smoking prevention and cessation as well as respiratory health in general to decrease COPD risk. Health care practitioners and

educators should communicate to different populations in culturally sensitive ways [59,60].

Educational materials and behavior change strategies may need to be customized according to different risk factors, beliefs, preferences, and technographics of different subpopulations [50,51]. On a basic level, people with lower levels of education or household income could be directed via phone geolocation to their local health care and to their community leaders for in-person guidance and support. Local leaders could then inform them about local smoking cessation programs and apps or websites that monitor air quality in their community. Trusted local authorities are helpful entry points in those communities, after which peer support and network effects spread the information.

AQI was not significant in the MLR, but it was significant in the GBT, albeit not as strongly as we expected. It could be that the AQI is more of a diffuse, macrolevel environmental factor that fluctuates over time, making some CBSAs worse on average, but with wide volatility, for example, as weather and wind directions change [61,62]. Therefore, AQI could have more of an indirect or interaction effect with other variables. Combining campaigns on smoking prevention with campaigns on air quality could create a holistic public health strategy, particularly—as our findings suggest—in communities considered vulnerable, that is, communities with lower education, higher unemployment, and lower household income. Subsidies for households in communities considered vulnerable to convert to more efficient, cleaner home heating and cooling methods would improve their home's air quality at a lower cost [63]. Research suggests that engaging communities in targeting their air quality issues can lead to more positive outcomes in both air quality and public health [64-66].

There is a small but growing body of research that uses ML models in health care and medicine. There is recognition that the models can be highly accurate, but there is no consensus yet on how to interpret the results in a way that meshes seamlessly with clinical practice. The following examples provide an overview.

Elshawi et al [67] compared model-agnostic explanations using 2 techniques, LIME and Shapley values, to interpret a ML model for predicting hypertension risk. LIME uses small subsets of the data, which may be idiosyncratic, to provide intuitive explanations, that is, rules. Shapley values are more theoretically sound and global, using all the available data, and are, therefore, less idiosyncratic than LIME, but they do not provide LIME's simple, linear explanations [67].

Hakkoum et al [68] conducted an extensive literature review of ML interpretability in medicine published between 1994 and 2020. The review found that there was no consensus on evaluation metrics or frameworks to assess the quality and utility of the interpretability methods [68]. The highest performing ML models did not translate easily into clinical rules.

Meng et al [69] reviewed the interpretability and fairness evaluation of deep learning models on MIMIC-IV data set, a large, publicly available benchmark for developing and evaluating the interpretability of high-performing ML models

that use sensitive demographic features. The review found that existing interpretation methods, for example, variable importance rankings, provide partial explanations without fully elucidating the model's complex decision logic.

In sum, there is no consensus on the best way to interpret high-performing ML models in health care. There are always trade-offs between accuracy and interpretability or explainability. We chose to use Shapley values because they represent the frontier in explainability, and they are similar to interpreting a multiple regression, interpreting 1 variable at a time, without the assumptions of linear models. In addition, Shapley values allow for nonlinear relationships between each independent (predictor) variable and the dependent variable. Variable importance plots in conjunction with Shapley values help us to identify the most important variables and characterize their relationships with COPD.

Our best MLR model had variance explained of 81.1%, MAE of 0.591, and SMAPE of 9.666. Our best ML model was the GBT, with variance explained of 85.7%, MAE of 0.456, and SMAPE of 6.956. The GBT explains most of the variance—4.6% more than the best MLR—with far less predictive error. The GBT's SMAPE (6.956) was 28% lower than that of the MLR's SMAPE (9.666). Similarly, the GBT's RMSE was 26% lower than the MLR's RMSE, and its MAE was 23% lower than that of the MLR. Real-world predictive accuracy should be similar to that found in the test data set because the test data were never used in the GBT's model development.

Our GBT performed strongly on the test data, with very little performance deterioration on the test data versus performance on the training and validation data. This demonstrates that the GBT model does not overfit the data. To interpret the GBT, we used a variable importance plot [34,70,71] and SHAP plots [72,73]. SHAP plots are useful for interpreting the strength of the pairwise relationships between predictor variable and COPD rate, showing the added nuances of the curvilinear plots. By doing so, we rendered transparent the “black-box model” [74-76], thus preserving interpretability and actionability, in addition to adding nonlinear nuance.

Limitations and Future Directions

This research has a few limitations. The data were obtained from 517 (56%) of the 929 CBSAs. We assumed that this was an adequate sample and that the remaining CBSAs that did not report the data were similar to those that did. Alternatively, it could be that the CBSA that did not report COPD rates did so because the rates were low, that is, COPD was not considered a major problem by the local public health officials. Data covering additional demographic variables, such as gender and age, in addition to occupational exposures and physical exercise, could be gathered [77-79]. Future research could develop separate models stratified by demographic variables such as race or ethnicity, assuming there are sufficient data for each categorical class. There could also be geopolitical variations in terms of population density as well as demographics, psychographics [80], and technographics [81,82].

Future data collection could focus on understanding racial or ethnic disparities. By collecting data more intensively from the minority populations, we could go deeper into understanding how their rates of COPD drop so dramatically. Is it related to active peer recommendations for better self-care in a predominantly White health care system and population? Is it related to successfully tailored smoking prevention or cessation programs? Data pertaining to answering these more specific questions could be collected to enhance our understanding of how best to tailor communications to different demographic or ethnographic groups.

All our models were structured as direct effects. We applied MLR and ML methods with data from CBSAs, which have significant variation in terms of health care access and quality. Using these models as a foundation, we should recognize the interconnectedness (ie, direct, indirect, and interactive) of pollutants and copollutants to fully understand COPD's complex etiology. Future research could model interaction, moderating, or mediating effects, perhaps with a structural equation model, to identify the direct and indirect effects of COPD, for example, showing how asthma may lead to COPD or to asthma-COPD overlap [77].

There are many research knowledge gaps in the health impacts of extreme air pollution, including the effects of interactions between temperature and air pollution on respiratory health due to climate change [83]. Future research directions could focus on modeling the direct and indirect links between environmental exposures and COPD. On the basis of those results, we could design interventions, such as air quality warning systems, to mitigate their impact. The findings would underscore the opportunities for public health regulations, public-private sector partnerships, private company entrepreneurship, and global initiatives to reduce environmental exposures.

Greenhouse gas emissions may exacerbate overall air quality [84-88], contributing indirectly to COPD. Future research could collect data on new, additional variables pertaining to climate change [89]. Wildfires, which are increasingly common, produce more carcinogens in the air, including high levels of particulate matter. This can directly decrease air quality or copollute with other ambient pollutants [90]. These problems have been shown to increase the odds of lung cancer [91], and it is plausible that they can also contribute to COPD.

The association between COPD and environmental pollutants, including tropospheric ozone, nitrogen dioxide, sulfur dioxide,

and occupational exposures, has been extensively investigated [8,91-94]. Coarse, fine, and ultrafine particulate matter have been studied extensively and linked to systemic oxidative stress, inflammation [95], atherosclerosis [96], and mortality [97] in the United States [98,99] and China [100-102]. Tropospheric ozone exposure by itself has been linked to impaired lung function and increased COPD-related hospital admissions [103-105]. Similarly, elevated levels of nitrogen dioxide and sulfur dioxide, which are common in cities and industrial work sites, have been linked to an increased risk of COPD in the general population [106,107] and older adults [108]. In sum, data pertaining to ambient pollution, for example, particulate matter, sulfur dioxide, and carbon monoxide, could be useful additional copollutant data to include in future models [6,86-88,91,109-111].

Conclusions

Our novel contributions in this paper include the following: (1) integration of multiple publicly available CDC data sources, (2) development of highly accurate models using linear and nonlinear methods, and (3) interpretation of the variable impacts for the best model. Smoking was the number 1 variable impacting the COPD rate, which was expected. Household income was the second most influential predictor variable. Four economic factors spanned the full range of influence, from large (household income) to moderate (education rate) to small (unemployment rate and GDP). The race or ethnicity variable also had a range of impacts, from moderately high (American Indian or Alaska Native percentage) to moderate (Black or Hispanic percentage) to small (White, Asian, or Native Hawaiian or Pacific Islander percentage).

This research demonstrates the power of ML methods in general and a GBT, which produced a highly accurate model of COPD rates. The computational complexity of a GBT enables it to obtain high accuracy, but health care policy makers may be reluctant to adopt it unless they can obtain a rule-based explanation. Furthermore, clinicians typically want to be able to explain, justify, and communicate results to others in an intuitive manner. Finally, there may be legal, auditing, or regulatory requirements concerning transparency. If the method is audited, and it cannot be clearly explained, there may be serious legal or financial consequences [72]. Consequently, it is important to have explainable models to open the "black box," rendering them interpretable and actionable [75,76]. This research shows that it is possible to do so.

Acknowledgments

The authors thank Malavika Andavilli, Xuhui Bai, Chulin Chen, Shan He, Lanxiang Shao, Moshe Shpits, and Ziyu Tang for contributions to an early version of this paper. The authors also thank the anonymous reviewers at the International Business School, Brandeis University, for their helpful comments.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Importance of model variables for gradient boosted tree.

[PNG File, 10 KB - [ai_v3ile58455_app1.png](#)]

References

1. Diaz-Guzman E, Khosravi M, Mannino DM. Asthma, chronic obstructive pulmonary disease, and mortality in the U.S. population. *COPD* 2011 Dec 08;8(6):400-407. [doi: [10.3109/15412555.2011.611200](#)] [Medline: [22149399](#)]
2. Hardin M, Silverman EK, Barr RG, Hansel NN, Schroeder JD, Make BJ, COPDGene Investigators. The clinical features of the overlap between COPD and asthma. *Respir Res* 2011 Sep 27;12(1):127 [FREE Full text] [doi: [10.1186/1465-9921-12-127](#)] [Medline: [21951550](#)]
3. Peden DB. Air pollution in asthma: effect of pollutants on airway inflammation. *Ann Allergy Asthma Immunol* 2001 Dec;87(6 Suppl 3):12-17. [doi: [10.1016/s1081-1206\(10\)62334-4](#)] [Medline: [11770676](#)]
4. Wong-Parodi G, Dias MB, Taylor M. Effect of using an indoor air quality sensor on perceptions of and behaviors toward air pollution (Pittsburgh empowerment library study): online survey and interviews. *JMIR Mhealth Uhealth* 2018 Mar 08;6(3):e48 [FREE Full text] [doi: [10.2196/mhealth.8273](#)] [Medline: [29519779](#)]
5. Iribarren SJ, Akande TO, Kamp KJ, Barry D, Kader YG, Suelzer E. Effectiveness of mobile apps to promote health and manage disease: systematic review and meta-analysis of randomized controlled trials. *JMIR Mhealth Uhealth* 2021 Jan 11;9(1):e21563 [FREE Full text] [doi: [10.2196/21563](#)] [Medline: [33427672](#)]
6. Valavanidis A, Vlachogianni T, Fiotakis K. Tobacco smoke: involvement of reactive oxygen species and stable free radicals in mechanisms of oxidative damage, carcinogenesis and synergistic effects with other respirable particles. *Int J Environ Res Public Health* 2009 Feb;6(2):445-462 [FREE Full text] [doi: [10.3390/ijerph6020445](#)] [Medline: [19440393](#)]
7. Berend N. Contribution of air pollution to COPD and small airway dysfunction. *Respirology* 2016 Feb 27;21(2):237-244 [FREE Full text] [doi: [10.1111/resp.12644](#)] [Medline: [26412571](#)]
8. Lissåker CT, Talbott EO, Kan H, Xu X. Status and determinants of individual actions to reduce health impacts of air pollution in US adults. *Arch Environ Occup Health* 2016 Dec 02;71(1):43-48. [doi: [10.1080/19338244.2014.988673](#)] [Medline: [25454076](#)]
9. Dransfield MT, Bailey WC. COPD: racial disparities in susceptibility, treatment, and outcomes. *Clin Chest Med* 2006 Sep;27(3):463-71, vii. [doi: [10.1016/j.ccm.2006.04.005](#)] [Medline: [16880056](#)]
10. Duru OK, Harawa NT, Kermah D, Norris KC. Allostatic load burden and racial disparities in mortality. *J Natl Med Assoc* 2012 Jan;104(1-2):89-95 [FREE Full text] [doi: [10.1016/s0027-9684\(15\)30120-6](#)] [Medline: [22708252](#)]
11. Alexander GR, Wingate MS, Bader D, Kogan MD. The increasing racial disparity in infant mortality rates: composition and contributors to recent US trends. *Am J Obstet Gynecol* 2008 Jan;198(1):51.e1-51.e9. [doi: [10.1016/j.ajog.2007.06.006](#)] [Medline: [17870043](#)]
12. Woolf SH, Johnson RE, Fryer GE, Rust G, Satcher D. The health impact of resolving racial disparities: an analysis of US mortality data. *Am J Public Health* 2004 Dec;94(12):2078-2081. [doi: [10.2105/ajph.94.12.2078](#)] [Medline: [15569956](#)]
13. Budd J, Miller BS, Manning EM, Lampos V, Zhuang M, Edelstein M, et al. Digital technologies in the public-health response to COVID-19. *Nat Med* 2020 Aug 07;26(8):1183-1192. [doi: [10.1038/s41591-020-1011-4](#)] [Medline: [32770165](#)]
14. Zicari RV, Brusseau J, Blomberg SN, Christensen HC, Coffee M, Ganapini MB, et al. On assessing trustworthy AI in healthcare. Machine learning as a supportive tool to recognize cardiac arrest in emergency calls. *Front Hum Dyn* 2021 Jul 8;3:7. [doi: [10.3389/fhumd.2021.673104](#)]
15. Wenstrup J, Havtorn JD, Borgholt L, Blomberg SN, Maaloe L, Sayre MR, et al. A retrospective study on machine learning-assisted stroke recognition for medical helpline calls. *NPJ Digit Med* 2023 Dec 19;6(1):235 [FREE Full text] [doi: [10.1038/s41746-023-00980-y](#)] [Medline: [38114611](#)]
16. Wu CT, Li GH, Huang CT, Cheng YC, Chen CH, Chien JY, et al. Acute exacerbation of a chronic obstructive pulmonary disease prediction system using wearable device data, machine learning, and deep learning: development and cohort study. *JMIR Mhealth Uhealth* 2021 May 06;9(5):e22591 [FREE Full text] [doi: [10.2196/22591](#)] [Medline: [33955840](#)]
17. Peng J, Chen C, Zhou M, Xie X, Zhou Y, Luo CH. A machine-learning approach to forecast aggravation risk in patients with acute exacerbation of chronic obstructive pulmonary disease with clinical indicators. *Sci Rep* 2020 Feb 20;10(1):3118 [FREE Full text] [doi: [10.1038/s41598-020-60042-1](#)] [Medline: [32080330](#)]
18. Moll M, Qiao D, Regan EA, Hunninghake GM, Make BJ, Tal-Singer R, et al. Machine learning and prediction of all-cause mortality in COPD. *Chest* 2020 Sep;158(3):952-964 [FREE Full text] [doi: [10.1016/j.chest.2020.02.079](#)] [Medline: [32353417](#)]
19. Lee Y, Kim E, Chae KJ, Lee CH. Machine learning predicts computed tomography (CT)-based normal regional lung function distribution in asthma and chronic obstructive pulmonary disease (COPD) (abstract). In: Proceedings of the American Thoracic Society 2023 International Conference. 2023 Presented at: ATSIC '23; May 19-24, 2023; Washington, DC URL: <https://tinyurl.com/2fwtcz4b> [doi: [10.1164/ajrccm-conference.2023.207.1_MeetingAbstracts.A4004](#)]
20. Estépar RS. Artificial intelligence in COPD: new venues to study a complex disease. *Barc Respir Netw Rev* 2020;6(2):144-160 [FREE Full text] [doi: [10.23866/BRNRev:2019-0014](#)] [Medline: [33521399](#)]
21. Fortis S, Gao Y, Baldomero AK, Sarrazin MV, Kaboli PJ. Association of rural living with COPD-related hospitalizations and deaths in US veterans. *Sci Rep* 2023 May 16;13(1):7887 [FREE Full text] [doi: [10.1038/s41598-023-34865-7](#)] [Medline: [37193770](#)]

22. Freimuth VS, Quinn SC, Thomas SB, Cole G, Zook E, Duncan T. African Americans' views on research and the Tuskegee syphilis study. *Soc Sci Med* 2001 Mar;52(5):797-808 [FREE Full text] [doi: [10.1016/s0277-9536\(00\)00178-7](https://doi.org/10.1016/s0277-9536(00)00178-7)] [Medline: [11218181](https://pubmed.ncbi.nlm.nih.gov/11218181/)]
23. Vozoris NT, Stanbrook MB. Smoking prevalence, behaviours, and cessation among individuals with COPD or asthma. *Respir Med* 2011 Mar;105(3):477-484 [FREE Full text] [doi: [10.1016/j.rmed.2010.08.011](https://doi.org/10.1016/j.rmed.2010.08.011)] [Medline: [20850288](https://pubmed.ncbi.nlm.nih.gov/20850288/)]
24. Ezzati M, Friedman AB, Kulkarni SC, Murray CJ. The reversal of fortunes: trends in county mortality and cross-county mortality disparities in the United States. *PLoS Med* 2008 Apr 22;5(4):e66 [FREE Full text] [doi: [10.1371/journal.pmed.0050066](https://doi.org/10.1371/journal.pmed.0050066)] [Medline: [18433290](https://pubmed.ncbi.nlm.nih.gov/18433290/)]
25. Housing patterns and core-based statistical areas. United States Census Bureau. URL: <https://www.census.gov/topics/housing/housing-patterns/about/core-based-statistical-areas.html> [accessed 2024-04-29]
26. NCSH provides timely and accurate health statistics for the United States. Centers for Disease Control and Prevention. URL: <https://www.cdc.gov/nchs/> [accessed 2024-08-05]
27. COVID-19 public data set. Centers for Disease Control and Prevention. URL: <https://chronicdata.cdc.gov/> [accessed 2024-08-05]
28. Chronic disease indicators (CDI). Centers for Disease Control and Prevention. URL: <https://www.cdc.gov/cdi/> [accessed 2024-08-06]
29. Office of public health data, surveillance, and technology (OPHDST). Centers for Disease Control and Prevention. 2024 Jan. URL: <https://www.cdc.gov/about/divisions-offices/ophdst.html> [accessed 2024-04-29]
30. Kaufman S, Rosset S, Perlich C, Stitelman O. Leakage in data mining: formulation, detection, and avoidance. *ACM Trans Knowl Discov Data* 2012 Dec 18;6(4):1-21. [doi: [10.1145/2382577.2382579](https://doi.org/10.1145/2382577.2382579)]
31. Hyndman RJ, Koehler AB. Another look at measures of forecast accuracy. *Int J Forecast* 2006 Oct;22(4):679-688. [doi: [10.1016/j.ijforecast.2006.03.001](https://doi.org/10.1016/j.ijforecast.2006.03.001)]
32. Willmott CJ, Matsuura K. Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Clim Res* 2005;30:79-82. [doi: [10.3354/cr030079](https://doi.org/10.3354/cr030079)]
33. Fayed HA, Atiya AF. Speed up grid-search for parameter selection of support vector machines. *Appl Soft Comput* 2019 Jul;80:202-210. [doi: [10.1016/j.asoc.2019.03.037](https://doi.org/10.1016/j.asoc.2019.03.037)]
34. Greenwell BM, Boehmke BC. Variable importance plots—an introduction to the VIP package. *R J* 2020;12(1):343-366. [doi: [10.32614/RJ-2020-013](https://doi.org/10.32614/RJ-2020-013)]
35. Antwarg L, Miller RM, Shapira B, Rokach L. Explaining anomalies detected by autoencoders using Shapley additive explanations. *Expert Syst Appl* 2021 Dec;186:115736 [FREE Full text] [doi: [10.1016/j.eswa.2021.115736](https://doi.org/10.1016/j.eswa.2021.115736)]
36. Garreau D, von Luxburg U. Explaining the explainer: a first theoretical analysis of LIME. *arXiv Preprint posted online January 10, 2020*. [FREE Full text] [doi: [10.1016/b978-0-32-396098-4.00020-x](https://doi.org/10.1016/b978-0-32-396098-4.00020-x)]
37. Gramegna A, Giudici P. SHAP and LIME: an evaluation of discriminative power in credit risk. *Front Artif Intell* 2021;4:752558 [FREE Full text] [doi: [10.3389/frai.2021.752558](https://doi.org/10.3389/frai.2021.752558)] [Medline: [34604738](https://pubmed.ncbi.nlm.nih.gov/34604738/)]
38. Hetlevik Ø, Melbye H, Gjesdal S. GP utilisation by education level among adults with COPD or asthma: a cross-sectional register-based study. *NPJ Prim Care Respir Med* 2016 Jun 09;26:16027 [FREE Full text] [doi: [10.1038/npjpcrm.2016.27](https://doi.org/10.1038/npjpcrm.2016.27)] [Medline: [27279354](https://pubmed.ncbi.nlm.nih.gov/27279354/)]
39. DiBonaventura MD, Paulose-Ram R, Su J, McDonald M, Zou KH, Wagner JS, et al. The impact of COPD on quality of life, productivity loss, and resource use among the elderly United States workforce. *COPD* 2012 Feb;9(1):46-57. [doi: [10.3109/15412555.2011.634863](https://doi.org/10.3109/15412555.2011.634863)] [Medline: [22292597](https://pubmed.ncbi.nlm.nih.gov/22292597/)]
40. Ford ES, Croft JB, Mannino DM, Wheaton AG, Zhang X, Giles WH. COPD surveillance--United States, 1999-2011. *Chest* 2013 Jul;144(1):284-305 [FREE Full text] [doi: [10.1378/chest.13-0809](https://doi.org/10.1378/chest.13-0809)] [Medline: [23619732](https://pubmed.ncbi.nlm.nih.gov/23619732/)]
41. Tillet T, Dillon C, Paulose-Ram R, Hnizdo E, Doney B. Estimating the U.S. prevalence of chronic obstructive pulmonary disease using pre- and post-bronchodilator spirometry: the National Health and Nutrition Examination Survey (NHANES) 2007-2010. *Respir Res* 2013 Oct 09;14(1):103 [FREE Full text] [doi: [10.1186/1465-9921-14-103](https://doi.org/10.1186/1465-9921-14-103)] [Medline: [24107140](https://pubmed.ncbi.nlm.nih.gov/24107140/)]
42. Zhang X, Holt JB, Lu H, Wheaton AG, Ford ES, Greenlund KJ, et al. Multilevel regression and poststratification for small-area estimation of population health outcomes: a case study of chronic obstructive pulmonary disease prevalence using the behavioral risk factor surveillance system. *Am J Epidemiol* 2014 Apr 15;179(8):1025-1033. [doi: [10.1093/aje/kwu018](https://doi.org/10.1093/aje/kwu018)] [Medline: [24598867](https://pubmed.ncbi.nlm.nih.gov/24598867/)]
43. Wheaton AG, Cunningham TJ, Ford ES, Croft JB, Centers for Disease Control and Prevention (CDC). Employment and activity limitations among adults with chronic obstructive pulmonary disease--United States, 2013. *MMWR Morb Mortal Wkly Rep* 2015 Mar 27;64(11):289-295 [FREE Full text] [Medline: [25811677](https://pubmed.ncbi.nlm.nih.gov/25811677/)]
44. Shiels MS, Chernyavskiy P, Anderson WF, Best AF, Haozous EA, Hartge P, et al. Trends in premature mortality in the USA by sex, race, and ethnicity from 1999 to 2014: an analysis of death certificate data. *Lancet* 2017 Mar 11;389(10073):1043-1054 [FREE Full text] [doi: [10.1016/S0140-6736\(17\)30187-3](https://doi.org/10.1016/S0140-6736(17)30187-3)] [Medline: [28131493](https://pubmed.ncbi.nlm.nih.gov/28131493/)]
45. Wheaton AG, Liu Y, Croft JB, VanFrank B, Croxton TL, Punturieri A, et al. Chronic obstructive pulmonary disease and smoking status - United States, 2017. *MMWR Morb Mortal Wkly Rep* 2019 Jun 21;68(24):533-538 [FREE Full text] [doi: [10.15585/mmwr.mm6824a1](https://doi.org/10.15585/mmwr.mm6824a1)] [Medline: [31220055](https://pubmed.ncbi.nlm.nih.gov/31220055/)]

46. Terzikhan N, Verhamme KM, Hofman A, Stricker BH, Brusselle GG, Lahousse L. Prevalence and incidence of COPD in smokers and non-smokers: the Rotterdam study. *Eur J Epidemiol* 2016 Aug;31(8):785-792 [FREE Full text] [doi: [10.1007/s10654-016-0132-z](https://doi.org/10.1007/s10654-016-0132-z)] [Medline: [26946425](https://pubmed.ncbi.nlm.nih.gov/26946425/)]
47. Doo JH, Kim SM, Park YJ, Kim KH, Oh YH, Kim JS, et al. Smoking cessation after diagnosis of COPD is associated with lower all-cause and cause-specific mortality: a nationwide population-based cohort study of South Korean men. *BMC Pulm Med* 2023 Jul 03;23(1):237 [FREE Full text] [doi: [10.1186/s12890-023-02533-1](https://doi.org/10.1186/s12890-023-02533-1)] [Medline: [37394482](https://pubmed.ncbi.nlm.nih.gov/37394482/)]
48. Perelman J, Alves J, Pfoertner TK, Moor I, Federico B, Kuipers MA, et al. The association between personal income and smoking among adolescents: a study in six European cities. *Addiction* 2017 Dec;112(12):2248-2256 [FREE Full text] [doi: [10.1111/add.13930](https://doi.org/10.1111/add.13930)] [Medline: [28667824](https://pubmed.ncbi.nlm.nih.gov/28667824/)]
49. COPD causes and risk factors. American Lung Association. 2024. URL: <https://tinyurl.com/tz5mdhbs> [accessed 2024-04-29]
50. Clausen A, Christensen ER, Jakobsen PR, Søndergaard J, Abrahamsen B, Rubin KH. Digital solutions for decision support in general practice - a rapid review focused on systems developed for the universal healthcare setting in Denmark. *BMC Prim Care* 2023 Dec 14;24(1):276 [FREE Full text] [doi: [10.1186/s12875-023-02234-y](https://doi.org/10.1186/s12875-023-02234-y)] [Medline: [38097998](https://pubmed.ncbi.nlm.nih.gov/38097998/)]
51. Eriksen J, Ebbesen M, Eriksen KT, Hjermitsev C, Knudsen C, Bertelsen P, et al. Equity in digital healthcare - the case of Denmark. *Front Public Health* 2023 Sep 6;11:1225222 [FREE Full text] [doi: [10.3389/fpubh.2023.1225222](https://doi.org/10.3389/fpubh.2023.1225222)] [Medline: [37744503](https://pubmed.ncbi.nlm.nih.gov/37744503/)]
52. Joo JY, Liu MF. Culturally tailored interventions for ethnic minorities: a scoping review. *Nurs Open* 2021 Sep 09;8(5):2078-2090 [FREE Full text] [doi: [10.1002/nop.2.733](https://doi.org/10.1002/nop.2.733)] [Medline: [34388862](https://pubmed.ncbi.nlm.nih.gov/34388862/)]
53. Radu I, Scheermesser M, Spiess MR, Schulze C, Händler-Schuster D, Pehlke-Milde J. Digital health for migrants, ethnic and cultural minorities and the role of participatory development: a scoping review. *Int J Environ Res Public Health* 2023 Oct 23;20(20):6962 [FREE Full text] [doi: [10.3390/ijerph20206962](https://doi.org/10.3390/ijerph20206962)] [Medline: [37887700](https://pubmed.ncbi.nlm.nih.gov/37887700/)]
54. Sood A, Petersen H, Liu C, Myers O, Shore XW, Gore BA, et al. Racial and ethnic minorities have a lower prevalence of airflow obstruction than non-Hispanic whites. *COPD* 2022;19(1):61-68 [FREE Full text] [doi: [10.1080/15412555.2022.2029384](https://doi.org/10.1080/15412555.2022.2029384)] [Medline: [35099333](https://pubmed.ncbi.nlm.nih.gov/35099333/)]
55. Gilkes A, Hull S, Durbaba S, Schofield P, Ashworth M, Mathur R, et al. Ethnic differences in smoking intensity and COPD risk: an observational study in primary care. *NPJ Prim Care Respir Med* 2017 Sep 04;27(1):50 [FREE Full text] [doi: [10.1038/s41533-017-0052-8](https://doi.org/10.1038/s41533-017-0052-8)] [Medline: [28871087](https://pubmed.ncbi.nlm.nih.gov/28871087/)]
56. Hall WJ, Chapman MV, Lee KM, Merino YM, Thomas TW, Payne BK, et al. Implicit racial/ethnic bias among health care professionals and its influence on health care outcomes: a systematic review. *Am J Public Health* 2015 Dec;105(12):e60-e76. [doi: [10.2105/AJPH.2015.302903](https://doi.org/10.2105/AJPH.2015.302903)] [Medline: [26469668](https://pubmed.ncbi.nlm.nih.gov/26469668/)]
57. Dovidio JF, Penner LA, Albrecht TL, Norton WE, Gaertner SL, Shelton JN. Disparities and distrust: the implications of psychological processes for understanding racial disparities in health and health care. *Soc Sci Med* 2008 Aug;67(3):478-486. [doi: [10.1016/j.socscimed.2008.03.019](https://doi.org/10.1016/j.socscimed.2008.03.019)] [Medline: [18508171](https://pubmed.ncbi.nlm.nih.gov/18508171/)]
58. Williams DR, Sternthal M. Understanding racial-ethnic disparities in health: sociological contributions. *J Health Soc Behav* 2010;51 Suppl(Suppl):S15-S27 [FREE Full text] [doi: [10.1177/0022146510383838](https://doi.org/10.1177/0022146510383838)] [Medline: [20943580](https://pubmed.ncbi.nlm.nih.gov/20943580/)]
59. Tucker CM, Marsiske M, Rice KG, Nielson JJ, Herman K. Patient-centered culturally sensitive health care: model testing and refinement. *Health Psychol* 2011 May;30(3):342-350 [FREE Full text] [doi: [10.1037/a0022967](https://doi.org/10.1037/a0022967)] [Medline: [21553978](https://pubmed.ncbi.nlm.nih.gov/21553978/)]
60. Chin JL. Culturally competent health care. *Public Health Rep* 2000;115(1):25-33 [FREE Full text] [doi: [10.1093/phr/115.1.25](https://doi.org/10.1093/phr/115.1.25)] [Medline: [10968582](https://pubmed.ncbi.nlm.nih.gov/10968582/)]
61. Berkowicz R, Palmgren F, Hertel O, Vignati E. Using measurements of air pollution in streets for evaluation of urban air quality — meteorological analysis and model calculations. *Sci Total Environ* 1996 Oct;189-190:259-265. [doi: [10.1016/0048-9697\(96\)05217-5](https://doi.org/10.1016/0048-9697(96)05217-5)]
62. Vardoulakis S, Fisher BE, Pericleous K, Gonzalez-Flesca N. Modelling air quality in street canyons: a review. *Atmos Environ* 2003 Jan;37(2):155-182. [doi: [10.1016/S1352-2310\(02\)00857-9](https://doi.org/10.1016/S1352-2310(02)00857-9)]
63. Jonidi Jafari A, Charkhloo E, Pasalari H. Urban air pollution control policies and strategies: a systematic review. *J Environ Health Sci Eng* 2021 Dec 08;19(2):1911-1940 [FREE Full text] [doi: [10.1007/s40201-021-00744-4](https://doi.org/10.1007/s40201-021-00744-4)] [Medline: [34900316](https://pubmed.ncbi.nlm.nih.gov/34900316/)]
64. Ward F, Lowther-Payne HJ, Halliday EC, Dooley K, Joseph N, Livesey R, et al. Engaging communities in addressing air quality: a scoping review. *Environ Health* 2022 Sep 19;21(1):89 [FREE Full text] [doi: [10.1186/s12940-022-00896-2](https://doi.org/10.1186/s12940-022-00896-2)] [Medline: [36117163](https://pubmed.ncbi.nlm.nih.gov/36117163/)]
65. Rosen LJ, Myers V, Winickoff JP, Kott J. Effectiveness of interventions to reduce tobacco smoke pollution in homes: a systematic review and meta-analysis. *Int J Environ Res Public Health* 2015 Dec 18;12(12):16043-16059 [FREE Full text] [doi: [10.3390/ijerph121215038](https://doi.org/10.3390/ijerph121215038)] [Medline: [26694440](https://pubmed.ncbi.nlm.nih.gov/26694440/)]
66. Titus AR, Kalousova L, Meza R, Levy DT, Thrasher JF, Elliott MR, et al. Smoke-free policies and smoking cessation in the United States, 2003-2015. *Int J Environ Res Public Health* 2019 Sep 02;16(17):3200 [FREE Full text] [doi: [10.3390/ijerph16173200](https://doi.org/10.3390/ijerph16173200)] [Medline: [31480698](https://pubmed.ncbi.nlm.nih.gov/31480698/)]
67. Elshawi R, Al-Mallah MH, Sakr S. On the interpretability of machine learning-based model for predicting hypertension. *BMC Med Inform Decis Mak* 2019 Jul 29;19(1):146 [FREE Full text] [doi: [10.1186/s12911-019-0874-0](https://doi.org/10.1186/s12911-019-0874-0)] [Medline: [31357998](https://pubmed.ncbi.nlm.nih.gov/31357998/)]

68. Hakkoum H, Abnane I, Idri A. Interpretability in the medical field: a systematic mapping and review study. *Applied Soft Computing* 2022 Mar;117:108391 [FREE Full text] [doi: [10.1016/j.asoc.2021.108391](https://doi.org/10.1016/j.asoc.2021.108391)]
69. Meng C, Trinh L, Xu N, Enouen J, Liu Y. Interpretability and fairness evaluation of deep learning models on MIMIC-IV dataset. *Sci Rep* 2022 May 03;12(1):7166 [FREE Full text] [doi: [10.1038/s41598-022-11012-2](https://doi.org/10.1038/s41598-022-11012-2)] [Medline: [35504931](https://pubmed.ncbi.nlm.nih.gov/35504931/)]
70. Genuer R, Poggi JM, Tuleau-Malot C. Variable selection using random forests. *Pattern Recognit Lett* 2010 Oct;31(14):2225-2236 [FREE Full text] [doi: [10.1016/j.patrec.2010.03.014](https://doi.org/10.1016/j.patrec.2010.03.014)]
71. Grömping U. Variable importance assessment in regression: linear regression versus random forest. *Am Stat* 2009 Nov;63(4):308-319. [doi: [10.1198/tast.2009.08199](https://doi.org/10.1198/tast.2009.08199)]
72. Adadi A, Berrada M. Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). *IEEE Access* 2018;6:52138-52160. [doi: [10.1109/access.2018.2870052](https://doi.org/10.1109/access.2018.2870052)]
73. Lundberg S, Lee SI. A unified approach to interpreting model predictions. In: *Proceedings of the 2017 Annual Conference on Neural Information Processing Systems*. 2017 Presented at: NIPS '17; December 4-9, 2017; Long Beach, CA URL: https://papers.nips.cc/paper_files/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html
74. Kamath SS, Ananthanarayana VS. Semantics-based web service classification using morphological analysis and ensemble learning techniques. *Int J Data Sci Anal* 2016 Oct 18;2(1-2):61-74. [doi: [10.1007/s41060-016-0026-x](https://doi.org/10.1007/s41060-016-0026-x)]
75. Gilpin LH, Bau D, Yuan BZ, Bajwa A, Specter M, Kagal L. Explaining explanations: an overview of interpretability of machine learning. In: *Proceedings of the 2018 IEEE 5th International Conference on Data Science and Advanced Analytics*. 2018 Presented at: DSAA '18; October 1-3, 2018; Turin, Italy p. 80-89 URL: <https://ieeexplore.ieee.org/document/8631448> [doi: [10.1109/dsaa.2018.00018](https://doi.org/10.1109/dsaa.2018.00018)]
76. Kottemann JE, Remus WE. A study of the relationship between decision model naturalness and performance. *MIS Q* 1989 Jun;13(2):171-181. [doi: [10.2307/248924](https://doi.org/10.2307/248924)]
77. Mannino DM. Chronic obstructive pulmonary disease: definition and epidemiology. *Respir Care* 2003 Dec;48(12):1185-1191 [FREE Full text] [Medline: [14651759](https://pubmed.ncbi.nlm.nih.gov/14651759/)]
78. Black C, Tesfaigzi Y, Bassein JA, Miller LA. Wildfire smoke exposure and human health: significant gaps in research for a growing public health issue. *Environ Toxicol Pharmacol* 2017 Oct;55:186-195 [FREE Full text] [doi: [10.1016/j.etap.2017.08.022](https://doi.org/10.1016/j.etap.2017.08.022)] [Medline: [28892756](https://pubmed.ncbi.nlm.nih.gov/28892756/)]
79. Kelley T, Kearney GD. Insights into the environmental health burden of childhood asthma. *Environ Health Insights* 2018;12:1178630218757445 [FREE Full text] [doi: [10.1177/1178630218757445](https://doi.org/10.1177/1178630218757445)] [Medline: [29497308](https://pubmed.ncbi.nlm.nih.gov/29497308/)]
80. Wells WD. Lifestyle and psychographics. In: Wells WD, editor. *Life Style and Psychographics: Definitions, Uses, and Problems*. New York, NY: Marketing Classics Press; 2011.
81. Assael H. A demographic and psychographic profile of heavy internet users and users by type of internet usage. *J Adv Res* 2005 Oct 12;45(01):93. [doi: [10.1017/S0021849905050014](https://doi.org/10.1017/S0021849905050014)]
82. Fox X, Purcell K. Chronic disease and the internet. Pew Research Center. URL: <https://www.pewresearch.org/internet/2010/03/24/chronic-disease-and-the-internet/> [accessed 2024-04-29]
83. Sujaritpong S, Dear K, Cope M, Walsh S, Kjellstrom T. Quantifying the health impacts of air pollution under a changing climate-a review of approaches and methodology. *Int J Biometeorol* 2014 Mar 25;58(2):149-160 [FREE Full text] [doi: [10.1007/s00484-012-0625-8](https://doi.org/10.1007/s00484-012-0625-8)] [Medline: [23354423](https://pubmed.ncbi.nlm.nih.gov/23354423/)]
84. Ramanathan V, Feng Y. Air pollution, greenhouse gases and climate change: global and regional perspectives. *Atmos Environ* 2009 Jan;43(1):37-50. [doi: [10.1016/j.atmosenv.2008.09.063](https://doi.org/10.1016/j.atmosenv.2008.09.063)]
85. Barbour E, Deakin EA. Smart growth planning for climate protection. *J Am Plann Assoc* 2012 Jan;78(1):70-86. [doi: [10.1080/01944363.2011.645272](https://doi.org/10.1080/01944363.2011.645272)]
86. Boyce JK, Pastor M. Clearing the air: incorporating air quality and environmental justice into climate policy. *Clim Change* 2013 Aug 6;120(4):801-814. [doi: [10.1007/S10584-013-0832-2](https://doi.org/10.1007/S10584-013-0832-2)]
87. Cushing L, Blaustein-Rejto D, Wander M, Pastor M, Sadd J, Zhu A, et al. Carbon trading, co-pollutants, and environmental equity: evidence from California's cap-and-trade program (2011-2015). *PLoS Med* 2018 Jul 10;15(7):e1002604 [FREE Full text] [doi: [10.1371/journal.pmed.1002604](https://doi.org/10.1371/journal.pmed.1002604)] [Medline: [29990353](https://pubmed.ncbi.nlm.nih.gov/29990353/)]
88. Kinney PL. Climate change, air quality, and human health. *Am J Prev Med* 2008 Nov;35(5):459-467. [doi: [10.1016/j.amepre.2008.08.025](https://doi.org/10.1016/j.amepre.2008.08.025)] [Medline: [18929972](https://pubmed.ncbi.nlm.nih.gov/18929972/)]
89. Strickland E, Andrew NG. Unbiggen AI: the AI pioneer says it's time for smart-sized, "data-centric" solutions to big issues. *IEEE Spectrum*. 2022. URL: <https://tinyurl.com/43bt77tn> [accessed 2024-04-29]
90. Liu JC, Pereira G, Uhl SA, Bravo MA, Bell ML. A systematic review of the physical health impacts from non-occupational exposure to wildfire smoke. *Environ Res* 2015 Jan;136:120-132 [FREE Full text] [doi: [10.1016/j.envres.2014.10.015](https://doi.org/10.1016/j.envres.2014.10.015)] [Medline: [25460628](https://pubmed.ncbi.nlm.nih.gov/25460628/)]
91. Kamis A, Cao R, He Y, Tian Y, Wu C. Predicting lung cancer in the United States: a multiple model examination of public health factors. *Int J Environ Res Public Health* 2021 Jun 06;18(11):6127 [FREE Full text] [doi: [10.3390/ijerph18116127](https://doi.org/10.3390/ijerph18116127)] [Medline: [34204140](https://pubmed.ncbi.nlm.nih.gov/34204140/)]
92. Andre M, Sartelet K, Moukhtar S, Andre JM, Redaelli M. Diesel, petrol or electric vehicles: what choices to improve urban air quality in the Ile-de-France region? A simulation platform and case study. *Atmos Environ* 2020 Nov;241:117752. [doi: [10.1016/j.atmosenv.2020.117752](https://doi.org/10.1016/j.atmosenv.2020.117752)]

93. Lam YF, Fu JS, Wu S, Mickley LJ. Impacts of future climate change and effects of biogenic emissions on surface ozone and particulate matter concentrations in the United States. *Atmos Chem Phys* 2011 May 23;11(10):4789-4806. [doi: [10.5194/acp-11-4789-2011](https://doi.org/10.5194/acp-11-4789-2011)]
94. Anenberg SC, Horowitz LW, Tong DQ, West JJ. An estimate of the global burden of anthropogenic ozone and fine particulate matter on premature human mortality using atmospheric modeling. *Environ Health Perspect* 2010 Sep;118(9):1189-1195 [FREE Full text] [doi: [10.1289/ehp.0901220](https://doi.org/10.1289/ehp.0901220)] [Medline: [20382579](https://pubmed.ncbi.nlm.nih.gov/20382579/)]
95. Pope CA, Hansen ML, Long RW, Nielsen KR, Eatough NL, Wilson WE, et al. Ambient particulate air pollution, heart rate variability, and blood markers of inflammation in a panel of elderly subjects. *Environ Health Perspect* 2004 Mar;112(3):339-345 [FREE Full text] [doi: [10.1289/ehp.6588](https://doi.org/10.1289/ehp.6588)] [Medline: [14998750](https://pubmed.ncbi.nlm.nih.gov/14998750/)]
96. Araujo JA. Particulate air pollution, systemic oxidative stress, inflammation, and atherosclerosis. *Air Qual Atmos Health* 2010 Nov 10;4(1):79-93 [FREE Full text] [doi: [10.1007/s11869-010-0101-8](https://doi.org/10.1007/s11869-010-0101-8)] [Medline: [21461032](https://pubmed.ncbi.nlm.nih.gov/21461032/)]
97. Dockery DW, Pope CA, Xu X, Spengler JD, Ware JH, Fay ME, et al. An association between air pollution and mortality in six U.S. cities. *N Engl J Med* 1993 Dec 09;329(24):1753-1759. [doi: [10.1056/NEJM199312093292401](https://doi.org/10.1056/NEJM199312093292401)] [Medline: [8179653](https://pubmed.ncbi.nlm.nih.gov/8179653/)]
98. Belleudi V, Faustini A, Stafoggia M, Cattani G, Marconi A, Perucci CA, et al. Impact of fine and ultrafine particles on emergency hospital admissions for cardiac and respiratory diseases. *Epidemiology* 2010 May;21(3):414-423. [doi: [10.1097/EDE.0b013e3181d5c021](https://doi.org/10.1097/EDE.0b013e3181d5c021)] [Medline: [20386174](https://pubmed.ncbi.nlm.nih.gov/20386174/)]
99. Schraufnagel DE. The health effects of ultrafine particles. *Exp Mol Med* 2020 Mar 17;52(3):311-317 [FREE Full text] [doi: [10.1038/s12276-020-0403-3](https://doi.org/10.1038/s12276-020-0403-3)] [Medline: [32203102](https://pubmed.ncbi.nlm.nih.gov/32203102/)]
100. Zhang Y, Ding Z, Xiang Q, Wang W, Huang L, Mao F. Short-term effects of ambient PM and PM air pollution on hospital admission for respiratory diseases: case-crossover evidence from Shenzhen, China. *Int J Hyg Environ Health* 2020 Mar;224:113418. [doi: [10.1016/j.ijheh.2019.11.001](https://doi.org/10.1016/j.ijheh.2019.11.001)] [Medline: [31753527](https://pubmed.ncbi.nlm.nih.gov/31753527/)]
101. Ni L, Chuang CC, Zuo L. Fine particulate matter in acute exacerbation of COPD. *Front Physiol* 2015 Oct 23;6:294 [FREE Full text] [doi: [10.3389/fphys.2015.00294](https://doi.org/10.3389/fphys.2015.00294)] [Medline: [26557095](https://pubmed.ncbi.nlm.nih.gov/26557095/)]
102. Li T, Hu R, Chen Z, Li Q, Huang S, Zhu Z, et al. Fine particulate matter (PM 2.5): the culprit for chronic lung diseases in China. *Chronic Dis Transl Med* 2018 Sep;4(3):176-186 [FREE Full text] [doi: [10.1016/j.cdtm.2018.07.002](https://doi.org/10.1016/j.cdtm.2018.07.002)] [Medline: [30276364](https://pubmed.ncbi.nlm.nih.gov/30276364/)]
103. Kim CS, Alexis NE, Rappold AG, Kehrl H, Hazucha MJ, Lay JC, et al. Lung function and inflammatory responses in healthy young adults exposed to 0.06 ppm ozone for 6.6 hours. *Am J Respir Crit Care Med* 2011 May 01;183(9):1215-1221 [FREE Full text] [doi: [10.1164/rccm.201011-1813OC](https://doi.org/10.1164/rccm.201011-1813OC)] [Medline: [21216881](https://pubmed.ncbi.nlm.nih.gov/21216881/)]
104. Mudway IS, Kelly FJ. Ozone and the lung: a sensitive issue. *Mol Aspects Med* 2000;21(1-2):1-48. [doi: [10.1016/s0098-2997\(00\)00003-0](https://doi.org/10.1016/s0098-2997(00)00003-0)] [Medline: [10804262](https://pubmed.ncbi.nlm.nih.gov/10804262/)]
105. Uysal N, Schapira RM. Effects of ozone on lung function and lung diseases. *Curr Opin Pulm Med* 2003 Mar;9(2):144-150. [doi: [10.1097/00063198-200303000-00009](https://doi.org/10.1097/00063198-200303000-00009)] [Medline: [12574695](https://pubmed.ncbi.nlm.nih.gov/12574695/)]
106. Hendryx M, Luo J, Chojenta C, Byles JE. Air pollution exposures from multiple point sources and risk of incident chronic obstructive pulmonary disease (COPD) and asthma. *Environ Res* 2019 Dec;179(Pt A):108783. [doi: [10.1016/j.envres.2019.108783](https://doi.org/10.1016/j.envres.2019.108783)] [Medline: [31590000](https://pubmed.ncbi.nlm.nih.gov/31590000/)]
107. Chen TM, Kuschner WG, Gokhale J, Shofer S. Outdoor air pollution: nitrogen dioxide, sulfur dioxide, and carbon monoxide health effects. *Am J Med Sci* 2007 Apr;333(4):249-256. [doi: [10.1097/MAJ.0b013e31803b900f](https://doi.org/10.1097/MAJ.0b013e31803b900f)] [Medline: [17435420](https://pubmed.ncbi.nlm.nih.gov/17435420/)]
108. Gong Jr H, Linn WS, Clark KW, Anderson KR, Geller MD, Sioutas C. Respiratory responses to exposures with fine particulates and nitrogen dioxide in the elderly with and without COPD. *Inhal Toxicol* 2005 Mar 06;17(3):123-132. [doi: [10.1080/08958370590904481](https://doi.org/10.1080/08958370590904481)] [Medline: [15788373](https://pubmed.ncbi.nlm.nih.gov/15788373/)]
109. Lary DJ, Lary T, Sattler B. Using machine learning to estimate global PM2.5 for environmental health studies. *Environ Health Insights* 2015;9(Suppl 1):41-52 [FREE Full text] [doi: [10.4137/EHI.S15664](https://doi.org/10.4137/EHI.S15664)] [Medline: [26005352](https://pubmed.ncbi.nlm.nih.gov/26005352/)]
110. Tai AP, Mickley LJ, Jacob DJ, Leibensperger EM, Zhang L, Fisher JA, et al. Meteorological modes of variability for fine particulate matter (PM2.5) air quality in the United States: implications for PM2.5 sensitivity to climate change. *Atmos Chem Phys* 2012;12(6):3131-3145 [FREE Full text] [doi: [10.5194/acp-12-3131-2012](https://doi.org/10.5194/acp-12-3131-2012)]
111. Peterson GC, Hogrefe C, Corrigan AE, Neas LM, Mathur R, Rappold AG. Impact of reductions in emissions from major source sectors on fine particulate matter-related cardiovascular mortality. *Environ Health Perspect* 2020 Jan;128(1):017005. [doi: [10.1289/ehp5692](https://doi.org/10.1289/ehp5692)]

Abbreviations

- AQI:** air quality index
- CBSA:** core-based statistical area
- CDC:** Centers for Disease Control and Prevention
- COPD:** chronic obstructive pulmonary disease
- GBT:** gradient boosted tree
- GDP:** gross domestic product

LIME: local interpretable model-agnostic explanations
MAE: mean absolute error
ML: machine learning
MLR: multiple linear regression
RMSE: root mean square error
SHAP: Shapley additive explanations
SMAPE: symmetric mean absolute percentage error

Edited by K El Emam, B Malin; submitted 15.03.24; peer-reviewed by GK Gupta, A Wani, S Mao; comments to author 20.04.24; revised version received 09.07.24; accepted 10.07.24; published 29.08.24.

Please cite as:

Kamis A, Gadia N, Luo Z, Ng SX, Thumbar M

Obtaining the Most Accurate, Explainable Model for Predicting Chronic Obstructive Pulmonary Disease: Triangulation of Multiple Linear Regression and Machine Learning Methods

JMIR AI 2024;3:e58455

URL: <https://ai.jmir.org/2024/1/e58455>

doi: [10.2196/58455](https://doi.org/10.2196/58455)

PMID:

©Arnold Kamis, Nidhi Gadia, Zilin Luo, Shu Xin Ng, Mansi Thumbar. Originally published in JMIR AI (<https://ai.jmir.org>), 29.08.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR AI, is properly cited. The complete bibliographic information, a link to the original publication on <https://www.ai.jmir.org/>, as well as this copyright and license information must be included.

Original Paper

Online Health Search Via Multidimensional Information Quality Assessment Based on Deep Language Models: Algorithm Development and Validation

Boya Zhang¹, MSc; Nona Naderi², PhD; Rahul Mishra¹, PhD; Douglas Teodoro¹, PhD

¹Department of Radiology and Medical Informatics, University of Geneva, Geneva, Switzerland

²Department of Computer Science, Université Paris-Saclay, Centre national de la recherche scientifique, Laboratoire Interdisciplinaire des Sciences du Numérique, Orsay, France

Corresponding Author:

Boya Zhang, MSc
Department of Radiology and Medical Informatics
University of Geneva
9 Chemin des Mines
Geneva, 1202
Switzerland
Phone: 41 782331908
Email: boya.zhang@unige.ch

Abstract

Background: Widespread misinformation in web resources can lead to serious implications for individuals seeking health advice. Despite that, information retrieval models are often focused only on the query-document relevance dimension to rank results.

Objective: We investigate a multidimensional information quality retrieval model based on deep learning to enhance the effectiveness of online health care information search results.

Methods: In this study, we simulated online health information search scenarios with a topic set of 32 different health-related inquiries and a corpus containing 1 billion web documents from the April 2019 snapshot of Common Crawl. Using state-of-the-art pretrained language models, we assessed the quality of the retrieved documents according to their usefulness, supportiveness, and credibility dimensions for a given search query on 6030 human-annotated, query-document pairs. We evaluated this approach using transfer learning and more specific domain adaptation techniques.

Results: In the transfer learning setting, the usefulness model provided the largest distinction between help- and harm-compatible documents, with a difference of +5.6%, leading to a majority of helpful documents in the top 10 retrieved. The supportiveness model achieved the best harm compatibility (+2.4%), while the combination of usefulness, supportiveness, and credibility models achieved the largest distinction between help- and harm-compatibility on helpful topics (+16.9%). In the domain adaptation setting, the linear combination of different models showed robust performance, with help-harm compatibility above +4.4% for all dimensions and going as high as +6.8%.

Conclusions: These results suggest that integrating automatic ranking models created for specific information quality dimensions can increase the effectiveness of health-related information retrieval. Thus, our approach could be used to enhance searches made by individuals seeking online health information.

(JMIR AI 2024;3:e42630) doi:[10.2196/42630](https://doi.org/10.2196/42630)

KEYWORDS

health misinformation; information retrieval; deep learning; language model; transfer learning; infodemic

Introduction

In today's digital age, individuals with diverse information needs, medical knowledge, and linguistic skills [1] turn to the

web for health advice and to make treatment decisions [2]. The mixture of facts and rumors in online resources [3] makes it challenging for users to discern accurate content [4]. To provide high-quality resources and enable properly informed decision-making [5], information retrieval systems should

differentiate between accurate and misinforming content [6]. Nevertheless, search engines rank documents mainly by their relevance to the search query [7], neglecting several health information quality concerns. Moreover, despite attempts by some search engines to combat misinformation [8], they lack transparency in terms of the methodology used and performance evaluation.

Health misinformation is defined as health-related information that is inaccurate or misleading based on current scientific evidence [9,10]. Due to the lack of health literacy for nonprofessionals [11] and the rise of the infodemic phenomenon [12]—the rapid spread of both accurate and inaccurate information about a medical topic on the internet [13]—health misinformation has become increasingly prevalent online. Topics related to misinformation, such as “vaccine” or “the relationship between coronavirus and 5G” have gained scientific interest across social media platforms like Twitter and Instagram [14–16] and among various countries [17]. Thus, the development of new credibility-centered search methods and assessment measures is crucial to address the pressing challenges in health-related information retrieval [18].

In recent years, numerous approaches have been introduced in the literature to categorize and assess misinformation according to multiple dimensions. Hesse et al [19] proposed 7 dimensions of *truthfulness*, which include *correctness*, *neutrality*, *comprehensibility*, *precision*, *completeness*, *speaker trustworthiness*, and *informativeness*. On the other hand, van der Linden [20] categorized an infodemic into 3 key dimensions: *susceptibility*, *spread*, and *immunization*. Information retrieval shared tasks, such as the Text Retrieval Conference (TREC) and the Conference and Labs of the Evaluation Forum (CLEF), have also started evaluating quality-based systems for health corpora using multiple dimensions [21,22]. The CLEF eHealth Lab Series proposed a benchmark to evaluate models according to the *relevance*, *readability*, and *credibility* of the retrieved information [23]. The TREC Health Misinformation Track 2021 proposed further metrics of *usefulness*, *supportiveness*, and *credibility* [24]. These dimensions also appear in the TREC Health Misinformation Track 2019 as *relevancy*, *efficacy*, and *credibility*, respectively. Additionally, models by Solainayagi and Ponnusamy [25] and Li et al [26] incorporated similar dimensions, emphasizing source *reliability* and the *credibility* of statements. These metrics represent some of the initial efforts to quantitatively assess the effectiveness of information retrieval engines in sourcing high-quality information, marking a shift from the traditional query-document relevance paradigm [27,28]. Despite their variations, these information quality metrics focus on the following 3 main common topics: (1) *relevancy* (also called *usefulness* or *informativeness*) of the source to the search topic, (2) *correctness* (also called *supportiveness* or *efficacy*) of the information according to the search topic, and (3) *credibility* (also called *trustworthiness*) of the source.

Thanks to these open shared tasks, several significant methodologies have been developed to improve the search for higher-quality health information. Although classical bag-of-words-based methods outperform neural network approaches in detecting health-related misinformation when training data are limited [29], more advanced approaches are

needed for web content. Specifically, research has proven the effectiveness of a hybrid approach that integrates classical handcrafted features with deep learning [18]. Further to this, multistage ranking systems [30,31], which couple the system with a label prediction model or use T5 [32] to rerank Okapi Best Match 25 (BM25) results, have been proposed. Particularly, Lima et al [30] considered the stance of the search query and engaged 2 assessors for an interactive search, integrating a continuous active learning method [33]. This approach sets a baseline of human effort in separating helpful from harmful web content. Despite their success, these models often do not take into account the different information quality aspects in their design.

In this study, we aimed to investigate the impact of multidimensional ranking on improving the quality of retrieved health-related information. Due to its coverage of the main information quality dimensions used in the scientific literature, we followed the empirical approach proposed in the TREC 2021 challenge, which considers *usefulness*, *supportiveness*, and *credibility* metrics, to propose a multidimensional ranking model. Using deep learning-based pretrained language models [34] through transfer learning and domain adaption approaches, we categorized the retrieved web resources according to different information quality dimensions. Specialized quality-oriented ranks obtained by reranking components were then fused [32] to provide the final ranked list. In contrast to prior studies, our approach relied on the automatic detection of harmful (or inaccurate) claims and used a multidimensional information quality model to boost helpful resources.

The main contributions of this work are 3-fold. We propose a multidimensional ranking model based on transfer learning and showed that it achieves state-of-the-art in automatic (ie, when the query stance is not provided) quality-centered ranking evaluations. We investigated our approach in 2 learning settings—transfer learning (ie, without query relevance judgments) and domain adaptation (ie, with query relevance judgments from a different corpus)—and demonstrated that they are capable of identifying more helpful documents than harmful ones, obtaining +5% and +7% help and harm compatibility scores, respectively. Last, we investigated how the combination of models specialized in different information dimensions impacts the quality of the results, and our analysis suggests that multidimensional aspects are crucial for extracting high-quality information, especially for unhelpful topics.

Methods

In this section, we introduce our search model based on multidimensional information quality aspects. We first describe the evaluation benchmark. We then detail the implementation methodology and describe our evaluation experiments using transfer learning and domain adaptation strategies.

TREC Health Misinformation Track 2021 Benchmark

Benchmark Data Set

To evaluate our approach, we used the TREC Health Misinformation Track 2021 benchmark [35] organized by the National Institute of Standards and Technology (NIST) [36].

The TREC Health Misinformation Track 2021 benchmark simulates web searches for specific health issues and interventions against a collection of English web documents [37]. For each topic, the benchmark annotates the quality of the retrieved web documents using a pooling approach, in which the top retrieved documents by systems participating in the challenge are evaluated according to their usefulness, correctness, and credibility and subsequently labeled as helpful or harmful. In this context, helpful documents are defined as those supportive of helpful treatments or that try to dissuade the reader from using unhelpful treatments, while harmful documents encourage the use of unhelpful treatments or dissuade the reader from using helpful treatments [24]. See Table S1 in [Multimedia Appendix 1](#) for more detail on the annotation.

Health-Related Topics

A topic in the TREC Health Misinformation Track 2021 benchmark consists of a health issue, an intervention, a query

Table 1. Examples of helpful and unhelpful topics with query and description.

Number	Query	Description	Stance
106	vitamin b12 sun exposure vitiligo	Can vitamin b12 and sun exposure together help treat vitiligo?	Helpful
102	tepid sponge bath reduce fever children	Is a tepid sponge bath a good way to reduce fever in children?	Unhelpful

Web Corpus

We used the Colossal Clean Crawled Corpus (C4), a collection of English-language web documents sourced from the public Common Crawl web scrape [38]. The corpus comprises 1 billion English documents from the April 2019 snapshot. To illustrate

that connects the corresponding intervention to the health problem, and a description that resembles the web search question using natural language. NIST only provided assessments for 35 of the initial 50 topics. Among the assessed topics, 3 were further excluded due to the absence of harmful documents. Consequently, the benchmark consisted of 32 topics: 14 labeled as helpful and 18 labeled as unhelpful. For these queries, a total of 6030 query-document pairs were human-annotated according to different scales of usefulness, correctness, and credibility scores. A “helpful topic” refers to an intervention beneficial for treating a health issue, while an “unhelpful topic” indicates an ineffective intervention. The stance is supported by evidence from a credible source. [Table 1](#) presents examples of the queries and descriptions of helpful and unhelpful topics.

the contradictory nature of the web information within the corpus, in [Table 2](#), we present 2 documents relevant to topic 102: “tepid sponge bath reduce fever in children.” Although an article advises against the intervention (“Do Not Use Sponging to Reduce a Fever”), another article advises it could be a viable option (“Sponging is an option for high fevers”).

Table 2. Examples of useful but contradictory documents for Topic 102: “Is a tepid sponge bath a good way to reduce fever in children?”.

Article information	Article 1	Article 2
Doc ID	en.noclean.c4-train.07165-of-07168.96468	en.noclean.c4-train.00001-of-07168.126948
Time stamp	2019-04-25T18:00:17Z	2019-04-23T20:13:31Z
Text	[...] Do Not Use Sponging to Reduce a Fever. It is not recommended that you use sponging to reduce your child’s fever. There is no information that shows that sponging or tepid baths improve your child’s discomfort associated with a fever or an illness. Cool or cold water can cause shivering and increase your child’s temperature. Also, never add rubbing alcohol to the water. Rubbing alcohol can be absorbed into the skin or inhaled, causing serious problems such as a coma. [...]	[...] Sponging With Lukewarm Water: Note: Sponging is an option for high fevers, but not required. It is rarely needed. When to Use: Fever above 104° F (40° C) AND doesn’t come down with fever meds. Always give the fever medicine at least an hour to work before sponging. How to Sponge: Use lukewarm water (85 - 90° F) (29.4 - 32.2° C). Sponge for 20-30 minutes. If your child shivers or becomes cold, stop sponging. [...]
URL	https://patiented.solutions.aap.org/	https://childrensclinicofraceland.com/

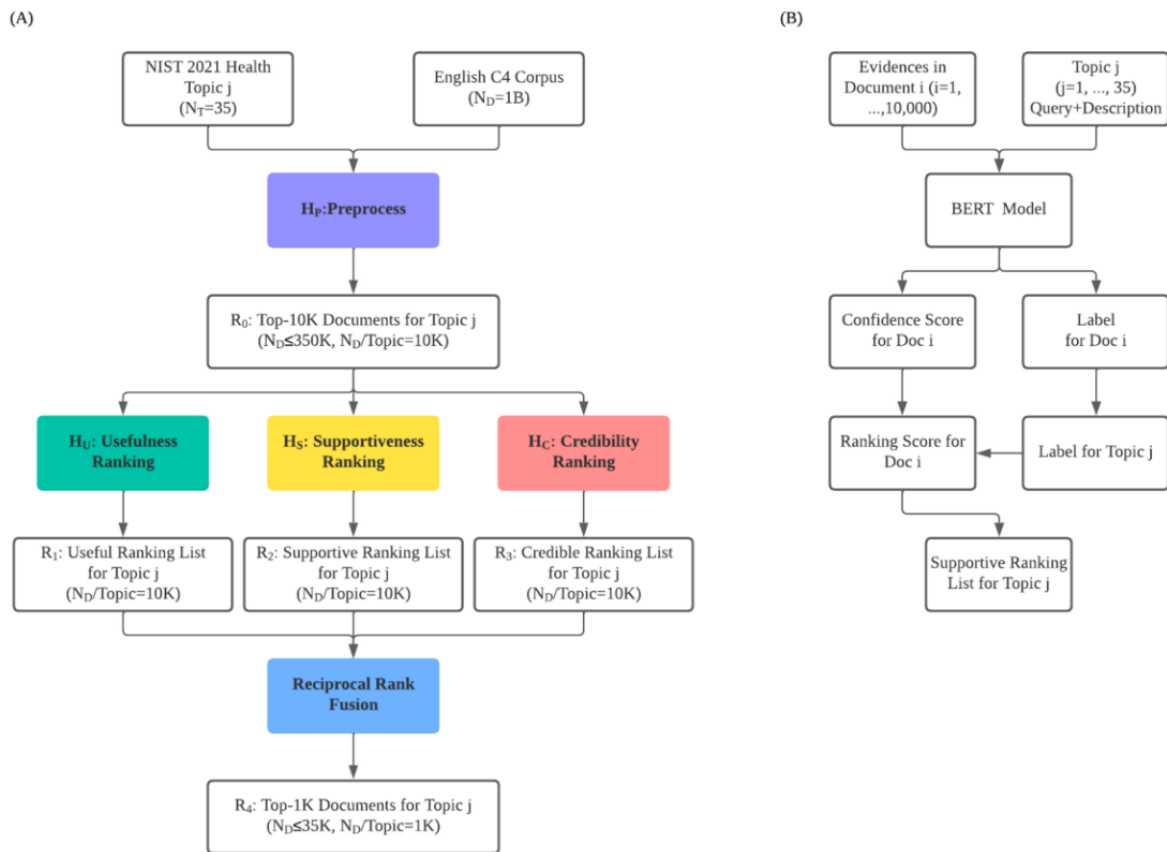
Quality-Based Multidimensional Ranking Conceptual Model

Phases

The quality-based multidimensional ranking model proposed in this work is presented in [Figure 1A](#). The information retrieval process can be divided into 2 phases: *preprocessing* and *multidimensional ranking*. In the preprocessing phase, for a

given topic j , N_D documents were retrieved based on their relevance (eg, using a BM25 model) [39]. In the multidimensional ranking phase, we further estimated the quality of the retrieved subset of documents according to the usefulness, supportiveness, and credibility dimensions. In the following sections, we describe the multidimensional ranking approach and its implementation using transfer learning and domain adaptation. We then describe the preprocessing step, which can be performed based on sparse or dense retrieval engines.

Figure 1. Quality-based multidimensional ranking models: (A) general pipeline, (B) supportiveness model for the transfer learning approach. BERT: Bidirectional Encoder Representations from Transformers; C4: Colossal Clean Crawled Corpus; NIST: National Institute of Standards and Technology.



Multidimensional Ranking

To provide higher-quality documents at the top ranks, we proposed using a set of machine learning models trained to classify documents according to the usefulness, supportiveness, and credibility dimensions. For the initial rank list obtained in the preprocessing phase (see details in the following sections), the documents were reranked in parallel according to the following strategies for usefulness, supportiveness, and credibility.

Usefulness

The usefulness dimension is defined as *the extent to which the document contains information that a search user would find useful in answering the topic's question*. In this sense, it defines how pertinent a document is to a given topic. Thus, to compute the usefulness of retrieved documents, topic-document similarity models based on pretrained language models, such as Bidirectional Encoder Representations from Transformers (BERT)-base [40], mono-BERT-large [41], and ELECTRA [42], could be used. Given a topic-document pair, the language model infers a score that gives the level of similarity between the 2 input text passages. Although bag-of-words models, such as BM25, provide a strong baseline for usefulness, they do not consider word relations by learning context-sensitive representations as is the case with the pretrained language models, which are used to enhance the quality of the original ranking [28].

Supportiveness

The supportiveness dimension defines whether *the document supports or dissuades the use of the treatment in the topic's question*. Therefore, it defines the stance of the document on the health topic. In this dimension, documents are identified under 3 levels: (1) supportive (ie, the document supports the treatment), (2) dissuasive (ie, the document refutes the treatment), and (3) neutral (ie, the document does not contain enough information to make the decision) [35]. To compute the supportiveness of a document to a given query, the system should be optimized so that documents that are either supportive, if the topic is helpful, or dissuasive, if the topic is unhelpful, are boosted to the top of the ranking list, which means that correct documents are boosted and misinforming documents are downgraded.

Credibility

The credibility dimension defines *whether the document is considered credible by the assessor*, that is, how trustworthy the source document is. To compute this dimension, the content of the document itself could be used (eg, leveraging language features, such as readability [43]), which is assessable using the Simple Measure of Gobbledygook index [44]. Moreover, document metadata could be also used, such as incoming and outgoing links, which can be calculated with link analysis algorithms [45], and URL addresses considered to be trusted sources [46].

Transfer Learning Implementation

To implement the multidimensional ranking model in scenarios in which relevance judgments are not available, we proposed multiple (pretrained) models for each of the quality dimensions using transfer learning.

Usefulness

In this reranking step, we created an ensemble of pretrained language models—BERT-base, mono-BERT-large, and ELECTRA—all fine-tuned in the MS MARCO [47] data set. Each model then predicted the similarity between the topic and the initial list of retrieved documents. Their results were finally combined using reciprocal rank fusion (RRF) [32].

Supportiveness

In this reranking step (Figure 1B), we created an ensemble of claim-checking models—robustly optimized BERT approach (RoBERTa)-Large [48], BioMedRoBERTa-base [49], and SciBERT-base [50]—which were fine-tuned on the FEVER [51] and SciFact [52] data sets. Claim-checking models take a claim and a document as the information source and validate the veracity of the claim based on the document content [53]. Most claim-checking models assume that document content is ground truth. Since this is not valid in the case of web documents, we added a further classification step that evaluates the correctness of the retrieved documents. We used the top-*k* assignments [44] provided by the claim-checking models to define whether the topic should be supported or refuted. The underlying assumption is that a scientific fact is defined by the largest number of evidence available for a topic. A higher rank is then given to the correct supportive or dissuasive documents, a medium rank is given to the neutral documents, and a lower rank is given to the incorrect supportive or dissuasive documents. The rank lists obtained for each model were then combined using RRF.

Credibility

In this step, we implemented a random forest classifier trained on the Microsoft Credibility data set [54] with a set of credibility-related features, such as readability, openpage rank [45], and the number of cascading style sheets (CSS). The data set manually rated 1000 web pages with credibility scores between 1 (“very noncredible”) and 5 (“very credible”). We converted these scores for a binary classification setting—that is, scores of 4 and 5 were considered as 1 or *credible*, and scores of 1, 2, and 3 were considered as 0 or *noncredible*. For the readability score, we relied on the Simple Measure of Gobbledygook index [44], which estimates the years of education an average person needs to understand a piece of writing. Following Schwarz and Morris [54], we retrieved a web page’s PageRank and used it as a feature to train the classifier. We further used the number of CSS style definitions to estimate the effort for the design of a web page [55]. Last, a list of credible websites scrapped from the Health On the Net search engine [46] for the evaluated topics was combined with the baseline model to explore better performance. The result of the classifier was added to the unitary value of the Health On the Net credible sites [46].

Domain Adaptation Implementation

To implement the multidimensional ranking model in scenarios in which relevance judgments are available, we compared different pretrained language models—BERT, BioBERT [56], and BigBird [57]—for each of the quality dimensions using domain adaptation. In this case, each model was fine-tuned to predict the relevance judgment of a specific dimension (ie, usefulness, supportiveness, and credibility). Although the input size was limited to 512 tokens for the first 2 models, BigBird allows up to 4096 tokens.

We used the TREC 2019 Decision Track [33] benchmark data set to fine-tune our specific quality dimension models. The TREC 2019 Decision Track benchmark data set contains 51 topics evaluated across 3 dimensions: relevance, effectiveness, and credibility. Adhering to the experimental design set by [58], we mapped the 2019 and 2021 benchmarks as follows. The relevance dimension (2019) was mapped to usefulness (2021), with highly relevant documents translated as very useful and relevant documents as useful. The effectiveness dimension (2019) was mapped to supportiveness (2021), with effective labels reinterpreted as supportive and ineffective as dissuasive. The credibility dimension (2019) was directly mapped to credibility (2021) using the same labels.

The 2019 track uses the ClueWeb12-B13 [59,60] corpus, which contains 50 million pages. More details on the TREC 2019 Decision Track [33] benchmark are provided in Table S2 in [Multimedia Appendix 1](#).

In the training phase, the language models received as input were the pair topic-document and a label for each dimension according to the 2019-2021 mapping strategy. At the inference time, given a topic-document pair from the TREC Health Misinformation Track 2021 benchmark, the model would infer its usefulness, supportiveness, or credibility based on the dimension on which it was trained.

Preprocessing or Ranking Phase

In the preprocessing step, which is initially executed to select a short list of candidate documents for the input query, a BM25 model was used. This step was performed using a bag-of-words model due to its efficiency. For the C4 snapshot collection, 2 indices were created, one using standard BM25 parameters and another fine-tuned using a collection of topics automatically generated (silver standard) from a set of 4985 indexed documents. For a given document, the silver topic was created based on the keyword2query [61] and doc2query [41] models to provide the query and description content, respectively. Using the silver topics and their respective documents, the BM25 parameters of the second index were then fine-tuned using grid search in a known-item search approach [62] (ie, for a given silver topic, the model should return in the top-1 the respective document used to generate it). The results of these 2 indices were fused using RRF.

Evaluation Metric

We followed the official TREC evaluation strategy and used the compatibility metric [46] to assess the performance of our models. Contrary to the classic information retrieval tasks, in

which the performance metric relies on the degree of relatedness between queries and documents, in quality retrieval, harmful documents should be penalized, especially if they are relevant to the query content. In this context, the compatibility metric calculates the similarity between the actual ranking R provided by a model and an ideal ranking I as provided by the query relevance annotations. According to Equation 1, the compatibility is calculated with the rank-biased overlap (RBO) [63] similarity metric, which is top-weighted, with greater weight placed at higher ranks to address the indeterminate and incomplete nature of web search results [64]:

where the parameter p represents the searcher's patience or persistence and is set to 0.95 in our experiments and K is the search depth and is set to 1000 to bring $pK-1$ as close to 0 as possible. As shown in Equation 2, an additional normalization step was added to accommodate short, truncated ideal results, so when there are fewer documents in the ideal ranking than in the actual ranking list, it does not influence the compatibility computation results:

To ensure that helpful and harmful documents are treated differently, even if both might be relevant to the query content, the assessments were divided into “help compatibility” (help) and “harm compatibility” (harm) metrics. To evaluate the ability of the system to separate helpful from harmful information, the “harm compatibility” results were then subtracted from the “help compatibility” results, which were marked as “help-harm compatibility” (help-harm). Overall, the more a ranking is compatible with the ideal helpful ranking, the better it is. Conversely, the more a ranking is compatible with the ideal harmful ranking, the worse it is.

Experimental Setup

The BM25 indices were created using the Elasticsearch framework (version 8.6.0). The number of documents N_D retrieved per topic in the preprocessing step was set to 10,000 in our experiments. The pretrained language models were based on open-source checkpoints from the HuggingFace platform [65] and were implemented using the open-source PyTorch framework. The language models used for the usefulness dimension and their respective HuggingFace implementations were BERT base (Capreolus/bert-base-msmarco), BERT large (castorini/monobert-large-msmarco-finetune-only), and ELECTRA (Capreolus/electra-base-msmarco). The language models used for the supportiveness dimension were RoBERTa base (allenai/biomed_roberta_base), RoBERTa large (roberta-large), and SciBERT (allenai/scibert_scivocab_uncased). For the credibility dimension, we used the random forest algorithm of the scikit-learn library. In the domain adaptation setup, we partitioned the 2019 labeled data set into training and validation sets using an 80%:20% split ratio; the latter was used to select the best models. We then fine-tuned BioBERT

(dmis-lab/biobert-base-cased-v1.1) with a batch size of 16, learning rate of 1^{-5} , and 20 epochs with early stopping set at 5 and utilizing the binary cross-entropy loss, which was optimized using the Adam optimizer. The BigBird model (google/bigbird-roberta-base) was fine-tuned with a batch size of 2, keeping all the other settings the same as the BioBERT model. All language models were fine-tuned using a single NVIDIA Tesla V100 graphics card with 32 GB of memory (see Multimedia Appendix 2 for more details). Results are reported using the compatibility and normalized discounted cumulative gain (nDCG) metrics. For reference, they were compared with the results of other participants of the official TREC Health Misinformation 2021 track, which have submitted runs for the automatic evaluation (ie, without using information about the topic stance). The code repository is available at [66].

Ethical Considerations

No human participants were involved in this research. All data used to build and evaluate the deep language models were publicly available and open access.

Results

Performance Results

In Table 3, we present the performance results of our quality-based retrieval models using the TREC Health Misinformation 2021 benchmark. Helpful compatibility (help) considers only helpful documents of the relevant judgment, while harmful compatibility (harm) considers only harmful documents and help-harm considers their compatibility difference (see Table S1 in Multimedia Appendix 1 for further detail). Additionally, we show the nDCG scores calculated using helpful (help) documents or harmful (harm) documents of the relevant judgment. The $helpful_T$, $unhelpful_T$, and all_T terms denote helpful topics, unhelpful topics, and all topics, respectively. H_U , H_S , and H_C rankings represent the combination of the preprocessing (H_p) results with the rerankings results for usefulness (H_U'), supportiveness (H_S'), and credibility (H_C'), respectively. For reference, we show our results compared with the models participating in the TREC Health Misinformation Track 2021: Pradeep et al [31] used the default BM25 ranker from Pyserini. Their reranking process incorporated a mix of mono and duo T5 models as well as Vera [67] on different topic fields. Abualsaud et al [68] created filtered collections that focus on filtering out nonmedical and unreliable documents, which were then used for retrieval with Anserini's BM25. Schlicht et al [69] also used Pyserini's BM25 ranker and Bio Sentence BERT to estimate usefulness and RoBERTa for credibility. The final score was a fusion of these individual rankings. Fernández-Pichel et al [70] used BM25 and RoBERTa for reranking and similarity assessment of the top 100 documents, trained an additional reliability classifier, and merged scores using CombSUM [71] or Borda Count. Bondarenko et al [72] used Anserini's BM25 and PyGaggle's MonoT5 for 2 baseline rankings, then reranked the top 20 from each using 3 argumentative axioms on seemingly argumentative queries.

Table 3. Performance results for the quality-based retrieval models.

Model	nDCG ^a		Compatibility				
	Help ^b ↑	Harm ^c ↓	Help ↑	Harm ↓	Help-harm ↑		
	all _T ^d	all _T	all _T	all _T	helpful _T ^e	unhelpful _T ^f	all
BM25 ^g [39]	0.516	0.360	0.122	0.144	0.158	-0.162	-0.022
Pradeep et al [31]	0.602	0.378	0.195 ^h	0.153	0.234 ^h	-0.106	0.043
Abualsaud et al [68]	0.302	0.185 ^h	0.164	0.123	0.179	-0.067	0.040
Schlicht et al [69]	0.438	0.309	0.121	0.103	0.157	-0.089	0.018
Fernández-Pichel et al [70]	0.603 ^h	0.363	0.163	0.155	0.163	-0.113	0.008
Bondarenko et al [72]	0.266	0.226	0.129	0.144	0.150	-0.144	-0.015
Transfer learning							
H_U ⁱ	0.538 ^j	0.324	0.142 ^j	0.087 ^h	0.156	-0.022 ^h	0.056 ^h
$H_U + H_S$ ^k	0.477	0.315 ^j	0.130	0.092	0.151	-0.049	0.038
$H_U + H_S + H_C$ ^l	0.484	0.320	0.137	0.095	0.169 ^j	-0.057	0.042
Domain adaptation							
H_U	0.510	0.327	0.128	0.100	0.146	-0.063	0.029
$H_U + H_S$	0.482	0.319	0.108	0.089	0.108	-0.050	0.019
$H_U + H_S + H_C$ ^l	0.502	0.325	0.131	0.094	0.147	-0.048	0.037

^anDCG: normalized discounted cumulative gain.

^bHelp: results considering only helpful documents in the relevance judgment.

^cHarm: results considering only harmful documents in the relevance judgment.

^dall_T: all topics.

^ehelpful_T: helpful topics.

^funhelpful_T: unhelpful topics.

^gBM25: Best Match 25.

^hBest performance.

ⁱ H_U : usefulness model.

^jBest performance among our models.

^k H_S : supportiveness model.

^l H_C : credibility model.

Our approach provides state-of-the-art results for automatic ranking systems in the transfer learning setting, with help-harm compatibility of +5.6%. This result was obtained with the usefulness model (H_U), which is the combination of preprocessing and usefulness reranking. It outperformed the default BM25 model [39] by 7% ($P=.04$) and the best automatic model from the TREC 2021 benchmark (Pradeep et al [31]) by 1%. In this case, although the help and harm compatibility metrics individually exhibited statistical significance ($P=.02$ and $P=.01$, respectively), the improvement in help-harm compatibility compared with the best automatic model was not statistically significant ($P=.70$). The usefulness model also stood out by achieving the best help and harm compatibility metrics among our models (14.2% and 8.7%, respectively; $P=.50$). Notice that, for the latter metric, the closest to 0, the better the performance. Interestingly, the usefulness model attained the

highest nDCG score on help for all topics as well ($P=.03$). The combination of usefulness, supportiveness, and credibility models ($H_U + H_S + H_C$) provided the best help-harm (+16.9%) for helpful topics among our models (H_U : $P=.40$; $H_U + H_S$: $P=.04$).

Meanwhile, when calculating nDCG scores on harm, the combination of usefulness and supportiveness model ($H_U + H_S$) in the transfer learning and domain adaptation settings outperformed the other model combinations ($P=.50$), indicating a different perspective of the best-performing model. Last, differently from what would be expected, in the domain adaptation setting, the performance was poorer than the simpler transfer learning approach (2% decrease on average for the compatibility metric; $P=.02$). See Table S4 in [Multimedia Appendix 3](#) for more information about using nDCG as a metric in a multidimensional evaluation.

Performance Stratification by Quality Dimension

In Table 4, we show the help, harm, and help-harm compatibility scores for the individual quality-based reranking models, which disregarded the preprocessing step (prime index). Additionally,

we provide the nDCG scores for a more comprehensive view of the models' performance. H_p represents the preprocessing, and H_U' , H_S' , and H_C' stand for rerankings for usefulness, supportiveness, and credibility, respectively.

Table 4. Performance results for the individual ranking models.

Setting and model	nDCG ^a		Compatibility				
	Help ^b ↑	Harm ^c ↓	Help ↑	Harm ↓	Help-harm ↑		
	all T ^d	all T	all T	all T	helpful T ^e	unhelpful T ^f	all T
H_p ^g	0.538 ^h	0.341	0.126 ^h	0.111	0.127 ^h	-0.072	0.015
Transfer learning							
H_U ^{i,j}	0.438	0.264	0.115	0.080	0.106	-0.020	0.036
H_S ^{j,k}	0.140	0.102 ^h	0.026	0.024	0.021	-0.013	0.002
H_C ^{j,l}	0.131	0.113	0.031	0.035	0.033	-0.032	-0.003
Domain adaptation							
H_U'	0.436	0.277	0.077	0.038	0.099	-0.008	0.039 ^h
H_S'	0.368	0.251	0.030	0.015 ^h	0.030	0.003 ^h	0.014
H_C'	0.443	0.296	0.079	0.064	0.104	-0.055	0.014

^anDCG: normalized discounted cumulative gain.

^bHelp: results considering only helpful documents in the relevance judgment.

^cHarm: results considering only harmful documents in the relevance judgment.

^dall T : all topics.

^ehelpful T : helpful topics.

^funhelpful T : unhelpful topics.

^g H_p : preprocess.

^hBest performance.

ⁱ H_U' : usefulness model.

^jUnlike H_U , H_S , and H_C , H_U' , H_S' , and H_C' rankings are not combined with H_p .

^k H_S' : supportiveness model.

^l H_C' : credibility model.

In the transfer learning setting, the usefulness model (H_U') achieved the highest help-harm compatibility (+3.6%; $P=.20$). The preprocessing model gave the best help compatibility (+12.7%; H_U' : $P=.70$; H_S' and H_C' : $P<.001$). Additionally, the preprocessing model yielded the highest nDCG score for help (H_U' : $P=.10$; H_S' and H_C' : $P<.001$). On the other hand, the preprocessing model showed the highest harm compatibility (+11.1%; H_U' : $P=.33$; H_S' and H_C' : $P<.01$). The combination of the preprocessing and usefulness models (ie, $H_U=+5.6\%$) improved the preprocessing model by 4.1% (from +1.5% to +5.6% on the help-harm compatibility; $P=.06$). For harm compatibility, the supportiveness model (H_S') achieved the best performance among the individual models (+2.4%; H_p : $P<.001$; H_U' : $P=.03$; H_C' : $P=.34$).

In the domain adaptation setting, the usefulness model (H_U') reached help-harm compatibility of +3.9%, similarly outperforming the other models ($P=.32$). The supportiveness

model (H_S') achieved the best performance on harm compatibility (+1.5%; $P=.07$) and on help-harm compatibility for unhelpful topics (+0.3%; $P=.50$). Notice that +0.3% is the only positive help-harm compatibility for harmful topics throughout all the individual and combined models on both settings including the preprocessing step. Last, in the domain adaption setting, the performance of individual models was better than the simpler transfer learning approach (1% increase on average for the compatibility metric; $P=.19$).

Reranking of the Top-N Documents

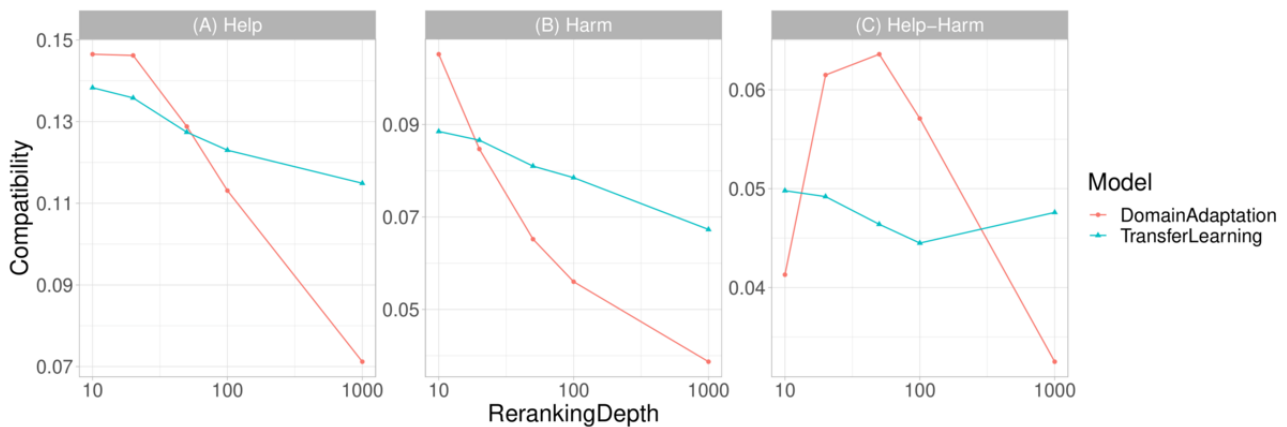
To further illustrate the effectiveness of the supportiveness and credibility dimensions, in Figure 2, we reranked only the top-n documents using the results of the usefulness model (H_U) as the basis. As we can see in Table 4, the overall effectiveness of the supportiveness (H_S') and credibility (H_C') models were considerably lower than that of the usefulness (H_U') model. The reason is that the relevance judgments were created using a

hierarchical approach: Only useful documents were further considered for supportiveness and credibility evaluations. As we reranked the documents in supportiveness and credibility dimensions without taking this hierarchy into account, their results might not be optimal. For example, low-ranking documents (ie, not useful) could have high credibility and, during the reranking process, could be boosted to the top ranks. Thus, we applied the supportiveness (H_S') and credibility (H_C') models to the usefulness model (H_U') results to rerank the top 10, 20, 50, 100, and 1000 documents, obtaining 2 new rankings, which were combined using RRF.

As the reranking depth increased from 10 to 1000, we observed a decrease in both help and harm compatibility. This suggests

that both helpful and harmful documents were downgraded due to the inclusion of less useful but potentially supportive or credible documents. In the transfer learning setting, as the reranking depth increased, the help-harm compatibility decreased until the depth reached 100. Beyond this point, we observed a slight increase at the depth of 1000. In the domain adaptation setting, the help-harm compatibility increased above +6% when the reranking depth was between 20 and 50. This implies that, following the procedure of human annotation, by considering only the more useful documents, the supportiveness and credibility dimensions can help retrieve more helpful than harmful documents.

Figure 2. Compatibility performance for the top 10, 20, 50, 100, and 1000 reranking depths taking the results of usefulness as the basis.



Quality Control

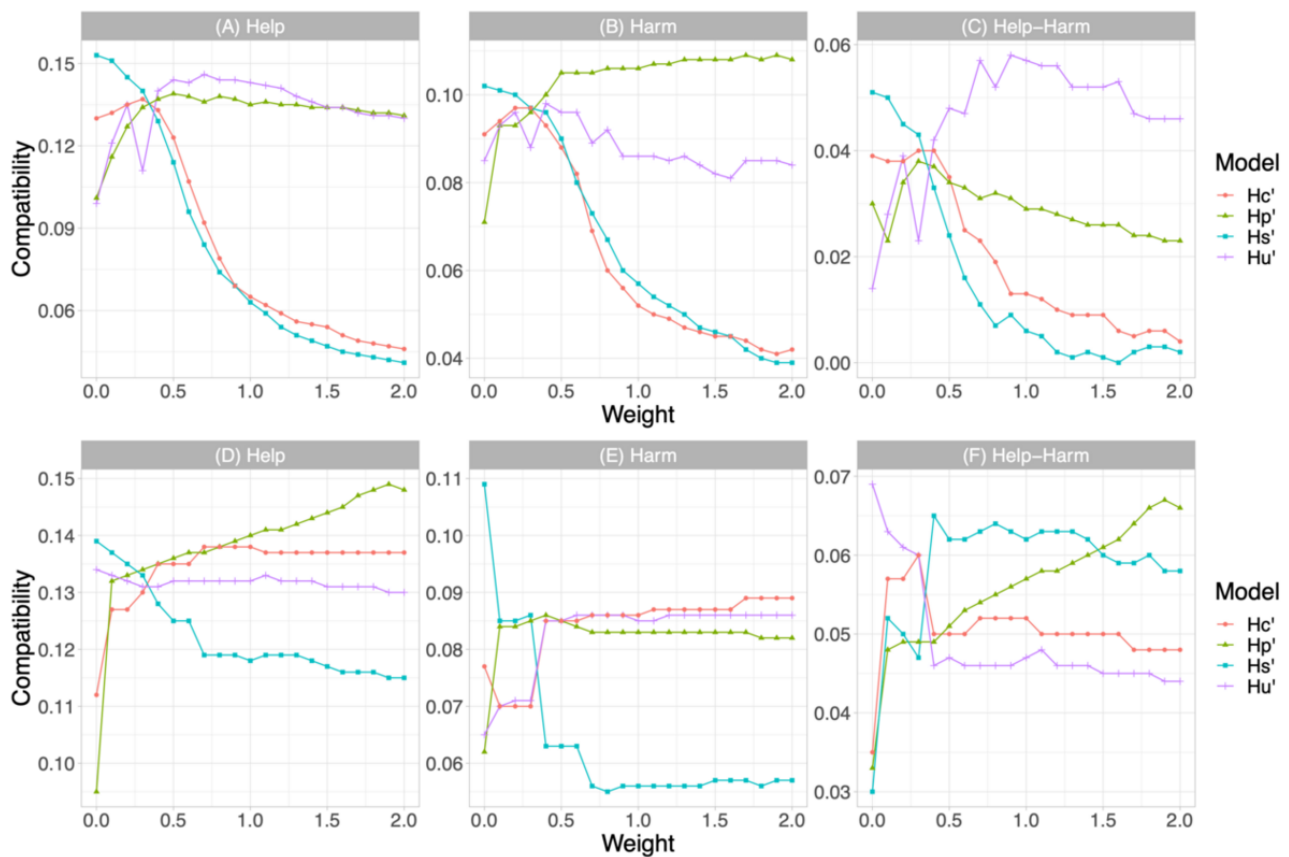
One of the advantages of the proposed multidimensional model is that we can optimize the results according to different quality metrics. In Figure 3, we show how the compatibility performance varies by changing the weight of the specific models (H_P , H_U' , H_S' , and H_C'). We normalized the score of the individual models to the unit and combined them linearly using a weight for 1 model between 0 and 2 while fixing the weight for the other 3 models at 0.33. For example, to see the influence of H_P in the final performance, we fixed the weights of H_U' , H_S' , and H_C' at 0.33 and varied the weight of H_P between 0 and 2. With weight 0, the reference model did not account for the final rank, while with weight 2, its impact was twice the sum of the other 3 models.

In the transfer learning setting, when we increased the weight of preprocessing and usefulness models, the help-harm compatibility increased to the best performance (+4.1% and +5.6%) then decreased slightly. For the supportiveness and

credibility dimensions, the help-harm compatibility began to decrease once the weight was added. These results imply that the compatibility decreases with the weight addition regardless of whether it is helpful compatibility, harmful compatibility, or the difference between the 2.

In the domain adaptation setting, when we increased the weight of preprocessing, supportiveness, and credibility models individually, the help-harm compatibility increased then converged to +6.6%, +5.9%, and +4.8%, respectively. For the usefulness model, the help-harm compatibility decreased once the weight was added until it converged to +4.4%. It is worth noticing that, by combining the rankings linearly, the help-harm compatibility obtained from the domain adaptation setting may exceed the results we obtained when performing ranking combination with RRF (+3.7%), as well as the state-of-the-art result (+5.6%) in the transfer learning setting. The highest help-harm compatibility scores for each weighting combination were +6.6%, +6.8%, +6.5%, and +5.9% when varying the weights of H_P , H_U' , H_S' , and H_C' , respectively.

Figure 3. Compatibility in the transfer learning approach (A-C) and compatibility in the domain adaptation approach (D-F), all with weights added to specific models.



Model Interpretation

To semantically explain the variation of help-harm compatibility, we set the search depth to 10. The help, harm, and help-harm compatibility of the 3 models are shown in Table 5. The help-harm compatibility was 1 when only helpful documents were retrieved in the top 10. Conversely, the help-harm compatibility was -1 when only harmful documents were retrieved in the top 10. A variation of 10% in the help or harm compatibility corresponded roughly to 1 helpful document exceeding the number of harmful documents retrieved in the top 10. Overall, the results show that retrieving relevant documents for health-related queries is hard, as, on average,

only 1.5 of 10 documents were relevant (helpful or harmful) to the topic. In addition, we interpreted that the 3 models retrieved, on average, twice the number of helpful documents as harmful documents. Particularly, H_U had, on average, around 1 more helpful than harmful document in the top 10, of the 1.5 relevant documents retrieved. We also present the same analysis results for the domain adaptation setting, which also implies that, when the rankings were combined with RRF, the transfer learning approach outperformed the domain adaptation approach. See more details about the average compatibility for all the topics as the search depth K varied in Figure S1 in Multimedia Appendix 3.

Table 5. Help, harm, and help-harm compatibility with search depth set to 10 for the transfer learning setting and domain adaptation setting.

Setting and model	Help ^a ↑	Harm ^b ↓	Help-harm ↑
Transfer learning			
H_U^c	0.112 ^d	0.047 ^d	0.065 ^d
$H_U + H_S^e$	0.088	0.050	0.038
$H_U + H_S + H_C^f$	0.099	0.056	0.044
Domain adaptation			
H_U	0.094	0.060	0.034
$H_U + H_S$	0.074	0.070	0.003
$H_U + H_S + H_C$	0.087	0.076	0.011

^aHelp: results considering only helpful documents in the relevance judgment.

^bHarm: results considering only harmful documents in the relevance judgment.

^c H_U : usefulness model.

^dBest performance.

^e H_S : supportiveness model.

^f H_C : credibility model.

Discussion

We propose a quality-based multidimensional ranking model to enhance the usefulness, supportiveness, and credibility of retrieved web resources for health-related queries. By adapting our approach in a transfer learning setting, we showed state-of-the-art results in the automatic quality ranking evaluation benchmark. We further explored the pipeline in a domain adaptation setting and showed that, in both settings, the proposed method can identify more helpful than harmful documents, as measured by +5% and +7% help-harm compatibility scores, respectively. By combining different reranking strategies, we showed that multidimensional aspects have a significant impact on retrieving high-quality information, particularly for unhelpful topics.

The quality of web documents is biased in terms of topic stance. For all models, helpful topics achieve higher help compatibility, while unhelpful topics achieve higher harm compatibility. The implication is that web documents centered around helpful topics are more likely to support the intervention and are helpful. On the other hand, web documents focusing on unhelpful topics present an equal chance of being supportive or dissuasive on the intervention and are helpful or harmful. Among other consequences, if web data are used to train large language models without meticulously crafted training examples using effective data set search methods [73], as the one proposed here, they are likely to further propagate health misinformation.

Automatic retrieval systems tend to find more helpful information on helpful topics with the information biased toward helpfulness and find more harmful information on unhelpful topics with the information slightly biased toward harmfulness. The help-harm compatibility ranged from +2.3% to +15.3% for helpful topics and from -5.7% to +0.2% for unhelpful topics. The difference shows that, for the improvement of quality-centered retrieval models, it is especially important to

focus on unhelpful topics. Moreover, although specialized models might provide enhanced effectiveness, their combination is not straightforward. In our experiments, we showed that supportiveness and credibility models should be applied only in the top 20 to 50 retrieved documents to achieve optimal performance.

Finding the correct stance automatically is another key component of the automatic model. Automatic models show the ability to prioritize helpful documents, resulting in positive help-harm compatibility. However, they are still far from state-of-the-art manual models, with help-harm compatibility scores ranging from +20.8% [68] to +25.9% [31]. We acknowledge that the help-harm compatibility can improve significantly with the correct stance given. This information is nevertheless unavailable in standard search environments; thus, the scenario analyzed in this work is more adapted to real-world applications.

This work has certain limitations. In the domain adaptation setting, we simplified the task to consider 2 classes within each dimension for the classification due to the limited variety available in the labeled data set. Alternatively, we could add other classes from documents that have been retrieved. Moreover, the number of topics used to evaluate our models was limited (n=32), despite including 6030 human-annotated, query-document pairs, and thus reflects only a small portion of misinformation use cases.

To conclude, the proliferation of health misinformation in web resources has led to mistrust and confusion among online health advice seekers. Automatic maintenance of factual discretion in web search results is the need of the hour. We propose a multidimensional information quality ranking model that utilizes usefulness, supportiveness, and credibility to strengthen the factual reliability of health advice search results. Experiments conducted on publicly available data sets show that the proposed model is promising, achieving state-of-the-art performance for

automatic ranking in comparison with various baselines implemented on the TREC Health Misinformation 2021 benchmark. Thus, the proposed approach could be used to improve online health searches and provide quality-enhanced

information for health information seekers. Future research could explore more granular classification models for each dimension, and a model simplification could provide an advantage for real-world implementations.

Acknowledgments

The study was funded by Innosuisse projects (funding numbers 55441.1 IP-ICT and 101.466 IP-ICT).

Data Availability

The data sets generated during and/or analyzed during this study are available in the Text Retrieval Conference (TREC) Health Misinformation Track repository [74] and GitLab repository [66].

Authors' Contributions

BZ, NN, and DT prepared the data, conceived and conducted the experiments, and analyzed the results. BZ, NN, and DT drafted the manuscript. All authors reviewed the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Additional Information on Benchmark Datasets.

[[PDF File \(Adobe PDF File\), 22 KB - ai_v3i1e42630_app1.pdf](#)]

Multimedia Appendix 2

Fine-Tuning in the Domain Adaptation Setting.

[[PDF File \(Adobe PDF File\), 69 KB - ai_v3i1e42630_app2.pdf](#)]

Multimedia Appendix 3

Supporting Experiment Results.

[[PDF File \(Adobe PDF File\), 195 KB - ai_v3i1e42630_app3.pdf](#)]

References

1. Goeuriot L, Jones GJF, Kelly L, Müller H, Zobel J. Medical information retrieval: introduction to the special issue. *Inf Retrieval J* 2016 Jan 11;19(1-2):1-5. [doi: [10.1007/s10791-015-9277-8](#)]
2. Chu JT, Wang MP, Shen C, Viswanath K, Lam TH, Chan SSC. How, when and why people seek health information online: qualitative study in Hong Kong. *Interact J Med Res* 2017 Dec 12;6(2):e24 [FREE Full text] [doi: [10.2196/ijmr.7000](#)] [Medline: [29233802](#)]
3. Lee JJ, Kang K, Wang MP, Zhao SZ, Wong JYH, O'Connor S, et al. Associations between COVID-19 misinformation exposure and belief with COVID-19 knowledge and preventive behaviors: cross-sectional online study. *J Med Internet Res* 2020 Nov 13;22(11):e22205 [FREE Full text] [doi: [10.2196/22205](#)] [Medline: [33048825](#)]
4. Ecker UKH, Lewandowsky S, Cook J, Schmid P, Fazio LK, Brashier N, et al. The psychological drivers of misinformation belief and its resistance to correction. *Nat Rev Psychol* 2022 Jan 12;1(1):13-29. [doi: [10.1038/s44159-021-00006-y](#)]
5. Krist AH, Tong ST, Aycock RA, Longo DR. Engaging patients in decision-making and behavior change to promote prevention. *Stud Health Technol Inform* 2017;240:284-302 [FREE Full text] [Medline: [28972524](#)]
6. Swire-Thompson B, Lazer D. Public health and online misinformation: challenges and recommendations. *Annu Rev Public Health* 2020 Apr 02;41(1):433-451 [FREE Full text] [doi: [10.1146/annurev-publhealth-040119-094127](#)] [Medline: [31874069](#)]
7. Sundin O, Lewandowski D, Haider J. Whose relevance? Web search engines as multisided relevance machines. *Asso for Info Science & Tech* 2021 Aug 21;73(5):637-642 [FREE Full text] [doi: [10.1002/asi.24570](#)]
8. Sullivan D. How Google delivers reliable information in Search. Google. 2020 Sep 10. URL: <https://blog.google/products/search/how-google-delivers-reliable-information-search/> [accessed 2024-04-18]
9. Di Sotto S, Viviani M. Health misinformation detection in the social web: an overview and a data science approach. *Int J Environ Res Public Health* 2022 Feb 15;19(4):A [FREE Full text] [doi: [10.3390/ijerph19042173](#)] [Medline: [35206359](#)]
10. Sylvia Chou W, Gaysynsky A, Cappella JN. Where we go from here: health misinformation on social media. *Am J Public Health* 2020 Oct;110(S3):S273-S275. [doi: [10.2105/ajph.2020.305905](#)]

11. Kickbusch I. Health literacy: addressing the health and education divide. *Health Promot Int* 2001 Sep;16(3):289-297. [doi: [10.1093/heapro/16.3.289](https://doi.org/10.1093/heapro/16.3.289)] [Medline: [11509466](https://pubmed.ncbi.nlm.nih.gov/11509466/)]
12. Suarez-Lledo V, Alvarez-Galvez J. Prevalence of health misinformation on social media: systematic review. *J Med Internet Res* 2021 Jan 20;23(1):e17187 [FREE Full text] [doi: [10.2196/17187](https://doi.org/10.2196/17187)] [Medline: [33470931](https://pubmed.ncbi.nlm.nih.gov/33470931/)]
13. Eysenbach G. How to fight an infodemic: the four pillars of infodemic management. *J Med Internet Res* 2020 Jun 29;22(6):e21820 [FREE Full text] [doi: [10.2196/21820](https://doi.org/10.2196/21820)] [Medline: [32589589](https://pubmed.ncbi.nlm.nih.gov/32589589/)]
14. Burki T. Vaccine misinformation and social media. *The Lancet Digital Health* 2019 Oct;1(6):e258-e259 [FREE Full text] [doi: [10.1016/s2589-7500\(19\)30136-0](https://doi.org/10.1016/s2589-7500(19)30136-0)]
15. Lotto M, Sá Menezes T, Zakir Hussain I, Tsao S, Ahmad Butt Z, P Morita P, et al. Characterization of false or misleading fluoride content on Instagram: infodemiology study. *J Med Internet Res* 2022 May 19;24(5):e37519 [FREE Full text] [doi: [10.2196/37519](https://doi.org/10.2196/37519)] [Medline: [35588055](https://pubmed.ncbi.nlm.nih.gov/35588055/)]
16. Mackey T, Purushothaman V, Haupt M, Nali M, Li J. Application of unsupervised machine learning to identify and characterise hydroxychloroquine misinformation on Twitter. *The Lancet Digital Health* 2021 Feb;3(2):e72-e75 [FREE Full text] [doi: [10.1016/s2589-7500\(20\)30318-6](https://doi.org/10.1016/s2589-7500(20)30318-6)]
17. Nsoesie EO, Cesare N, Müller M, Ozonoff A. COVID-19 misinformation spread in eight countries: exponential growth modeling study. *J Med Internet Res* 2020 Dec 15;22(12):e24425 [FREE Full text] [doi: [10.2196/24425](https://doi.org/10.2196/24425)] [Medline: [33264102](https://pubmed.ncbi.nlm.nih.gov/33264102/)]
18. Upadhyay R, Pasi G, Viviani M. Health misinformation detection in web content: a structural-, content-based, and context-aware approach based on Web2Vec. *GoodIT '21: Proceedings of the Conference on Information Technology for Social Good* 2021 Sep:19-24. [doi: [10.1145/3462203.3475898](https://doi.org/10.1145/3462203.3475898)]
19. Hesse BW, Nelson DE, Kreps GL, Croyle RT, Arora NK, Rimer BK, et al. Trust and sources of health information: the impact of the Internet and its implications for health care providers: findings from the first Health Information National Trends Survey. *Arch Intern Med* 2005;165(22):2618-2624. [doi: [10.1001/archinte.165.22.2618](https://doi.org/10.1001/archinte.165.22.2618)] [Medline: [16344419](https://pubmed.ncbi.nlm.nih.gov/16344419/)]
20. van der Linden S. Misinformation: susceptibility, spread, and interventions to immunize the public. *Nat Med* 2022 Mar 10;28(3):460-467. [doi: [10.1038/s41591-022-01713-6](https://doi.org/10.1038/s41591-022-01713-6)] [Medline: [35273402](https://pubmed.ncbi.nlm.nih.gov/35273402/)]
21. Pogacar FA, Ghenai A, Smucker MD, Clarke CLA. The positive and negative influence of search results on people's decisions about the efficacy of medical treatments. *ICTIR '17: Proceedings of the ACM SIGIR International Conference on Theory of Information Retrieval* 2017 Oct:209-216. [doi: [10.1145/3121050.3121074](https://doi.org/10.1145/3121050.3121074)]
22. Upadhyay R, Pasi G, Viviani M. An overview on evaluation labs and open issues in health-related credible information retrieval. *Proceedings of the 11th Italian Information Retrieval Workshop 2021* 2021:1 [FREE Full text]
23. Suominen H, Kelly L, Goeriot L, Krallinger M. CLEF eHealth Evaluation Lab 2020. *Advances in Information Retrieval* 2020;12036:587-594. [doi: [10.1007/978-3-030-45442-5_76](https://doi.org/10.1007/978-3-030-45442-5_76)]
24. Clarke CLA, Maistro M, Smucker MD. Overview of the TREC 2021 Health Misinformation Track. *NIST Special Publication: NIST SP 500-335: The Thirtieth Text REtrieval Conference (TREC 2021) Proceedings* 2022:1 [FREE Full text]
25. Solainayagi P, Ponnusamy R. Trustworthy media news content retrieval from web using truth content discovery algorithm. *Cognitive Systems Research* 2019 Aug;56:26-35. [doi: [10.1016/j.cogsys.2019.01.002](https://doi.org/10.1016/j.cogsys.2019.01.002)]
26. Li L, Qin B, Ren W, Liu T. Truth discovery with memory network. *Tsinghua Science and Technology* 2017 Dec;22(6):609-618. [doi: [10.23919/tst.2017.8195344](https://doi.org/10.23919/tst.2017.8195344)]
27. Zhang E, Gupta N, Tang R, Han X, Pradeep R, Lu K, et al. Covidex: neural ranking models and keyword search infrastructure for the COVID-19 open research dataset. *Proceedings of the First Workshop on Scholarly Document Processing* 2020:31-41. [doi: [10.18653/v1/2020.sdp-1.5](https://doi.org/10.18653/v1/2020.sdp-1.5)]
28. Teodoro D, Ferdowsi S, Borisssov N, Kashani E, Vicente Alvarez D, Copara J, et al. Information retrieval in an infodemic: the case of COVID-19 publications. *J Med Internet Res* 2021 Sep 17;23(9):e30161 [FREE Full text] [doi: [10.2196/30161](https://doi.org/10.2196/30161)] [Medline: [34375298](https://pubmed.ncbi.nlm.nih.gov/34375298/)]
29. Fernández-Pichel M, Losada DE, Pichel JC, Elswelier D. Comparing Traditional Neural Approaches for Detecting Health-Related Misinformation. In: Candan KS, editor. *Experimental IR Meets Multilinguality, Multimodality, and Interaction. CLEF 2021. Lecture Notes in Computer Science()*, vol 12880. Cham, Switzerland: Springer International Publishing; 2021:78-90.
30. Lima LC, Wright DB, Augenstein I, Maistro M. University of Copenhagen participation in TREC Health Misinformation Track 2020. *NIST Special Publication: NIST SP 1266: The Twenty-Ninth Text REtrieval Conference (TREC 2020) Proceedings* 2021:1 [FREE Full text]
31. Pradeep R, Ma X, Nogueira R, Lin J. Vera: prediction techniques for reducing harmful misinformation in consumer health search. *SIGIR '21: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval* 2021 Jul:2066-2070. [doi: [10.1145/3404835.3463120](https://doi.org/10.1145/3404835.3463120)]
32. Cormack GV, Clarke CLA, Buettcher S. Reciprocal rank fusion outperforms condorcet and individual rank learning methods. *SIGIR '09: Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval* 2009:758-759. [doi: [10.1145/1571941.1572114](https://doi.org/10.1145/1571941.1572114)]
33. Abualsaud M, Lioma C, Maistro M, Smucker M, Zuccon G. Overview of the TREC 2019 Decision Track. *NIST Special Publication: SP 500-331: The Twenty-Eighth Text REtrieval Conference (TREC 2019) Proceedings* 2020:1 [FREE Full text]

34. Zhang B, Naderi N, Jaume-Santero F, Teodoro D. DS4DH at TREC Health Misinformation 2021: multi-dimensional ranking models with transfer learning and rank fusion. NIST Special Publication: NIST SP 500-335: The Thirtieth Text REtrieval Conference (TREC 2021) Proceedings 2022:1 [[FREE Full text](#)]
35. Clarke CLA, Rizvi S, Smucker MD, Maistro M, Zuccon G. Overview of the TREC 2020 Health Misinformation Track. NIST Special Publication: NIST SP 1266: The Twenty-Ninth Text REtrieval Conference (TREC 2020) Proceedings 2021:1 [[FREE Full text](#)]
36. National Institute of Standards and Technology. URL: <https://www.nist.gov/> [accessed 2024-04-18]
37. Raffel C, Shazeer N, Roberts A, Lee K, Narang S, Matena M, et al. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research* 2020;21(140):1-67 [[FREE Full text](#)]
38. Common Crawl. URL: <https://commoncrawl.org/> [accessed 2024-04-18]
39. Robertson S, Zaragoza H. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval* 2009 Apr;3(4):333-389. [doi: [10.1561/1500000019](https://doi.org/10.1561/1500000019)]
40. Li C, Yates A, MacAvaney S, He B, Sun Y. PARADE: Passage Representation Aggregation for Document Reranking. *ACM Transactions on Information Systems* 2023 Sep 27;42(2):1-26. [doi: [10.1145/3600088](https://doi.org/10.1145/3600088)]
41. Nogueira R, Yang W, Cho K, Lin J. Multi-stage document ranking with BERT. arXiv Preprint posted online on October 31, 2019. [doi: [10.48550/arXiv.1910.14424](https://doi.org/10.48550/arXiv.1910.14424) [Focus to learn more](#)]
42. Clark K, Luong MH, Le QV, Manning CD. Electra: Pre-training text encoders as discriminators rather than generators. arXiv Preprint posted online on March 23, 2020. [doi: [10.48550/arXiv.2003.10555](https://doi.org/10.48550/arXiv.2003.10555)]
43. Zhou S, Jeong H, Green PA. How consistent are the best-known readability equations in estimating the readability of design standards? *IEEE Trans. Profess. Commun* 2017 Mar;60(1):97-111. [doi: [10.1109/tpc.2016.2635720](https://doi.org/10.1109/tpc.2016.2635720)]
44. Grabeel KL, Russomanno J, Oelschlegel S, Tester E, Heidel RE. Computerized versus hand-scored health literacy tools: a comparison of Simple Measure of Gobbledygook (SMOG) and Flesch-Kincaid in printed patient education materials. *J Med Libr Assoc* 2018 Jan 12;106(1):38-45 [[FREE Full text](#)] [doi: [10.5195/jmla.2018.262](https://doi.org/10.5195/jmla.2018.262)] [Medline: [29339932](https://pubmed.ncbi.nlm.nih.gov/29339932/)]
45. getPageRank. OpenPageRank. URL: <https://www.domcop.com/openpagerank/documentation> [accessed 2024-04-18]
46. Boyer C, Selby M, Scherrer J, Appel R. The Health On the Net Code of Conduct for medical and health websites. *Comput Biol Med* 1998 Sep;28(5):603-610 [[FREE Full text](#)] [doi: [10.1016/s0010-4825\(98\)00037-7](https://doi.org/10.1016/s0010-4825(98)00037-7)] [Medline: [9861515](https://pubmed.ncbi.nlm.nih.gov/9861515/)]
47. Bajaj P, Campos D, Craswell N, Deng L, Gao J, Liu X, et al. MS MARCO: a human generated machine reading comprehension dataset. arXiv Preprint posted online on October 31, 2018. [doi: [10.48550/arXiv.1611.09268](https://doi.org/10.48550/arXiv.1611.09268)]
48. Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, et al. RoBERTa: a robustly optimized BERT pretraining approach. arXiv Preprint posted online on July 26, 2019. [doi: [10.48550/arXiv.1907.11692](https://doi.org/10.48550/arXiv.1907.11692)]
49. Gururangan S, Marasović A, Swayamdipta S, Lo K, Beltagy I, Downey D, et al. Don't stop pretraining: adapt language models to domains and tasks. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* 2020:8342-8360 [[FREE Full text](#)] [doi: [10.18653/v1/2020.acl-main.740](https://doi.org/10.18653/v1/2020.acl-main.740)]
50. Beltagy I, Lo K, Cohan A. SciBERT: a pretrained language model for scientific text. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP) 2019*:3615-3620. [doi: [10.18653/v1/d19-1371](https://doi.org/10.18653/v1/d19-1371)]
51. Aly R, Guo Z, Schlichtkrull M, Thorne J, Vlachos A, Christodoulopoulos C, et al. The fact extraction and verification over unstructured and structured information (FEVEROUS) shared task. *Proceedings of the Fourth Workshop on Fact Extraction and VERification (FEVER) 2021*:1-13. [doi: [10.18653/v1/2021.fever-1.1](https://doi.org/10.18653/v1/2021.fever-1.1)]
52. Wadden D, Lin S, Lo K, Wang L, van Zuylen M, Cohan A, et al. Fact or fiction: verifying scientific claims. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP) 2020*:7534-7550. [doi: [10.18653/v1/2020.emnlp-main.609](https://doi.org/10.18653/v1/2020.emnlp-main.609)]
53. Stambach D, Zhang B, Ash E. The choice of textual knowledge base in automated claim checking. *Journal of Data and Information Quality* 2023;15(1):1-22. [doi: [10.1145/3561389](https://doi.org/10.1145/3561389)]
54. Schwarz J, Morris M. Augmenting web pages and search results to support credibility assessment. *CHI '11: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems 2011*:1245-1254. [doi: [10.1145/1978942.1979127](https://doi.org/10.1145/1978942.1979127)]
55. Olteanu A, Peshterliev S, Liu X, Aberer K. Web credibility: Features exploration and credibility prediction. In: Serdyukov P, Braslavski P, Kuznetsov SO, Kamps J, Rügger S, Agichtein E, et al, editors. *Advances in Information Retrieval. ECIR 2013. Lecture Notes in Computer Science, vol 7814. Berlin, Germany: Springer; 2013*:557-568.
56. Lee J, Yoon W, Kim S, Kim D, Kim S, So C, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* 2020 Feb 15;36(4):1234-1240 [[FREE Full text](#)] [doi: [10.1093/bioinformatics/btz682](https://doi.org/10.1093/bioinformatics/btz682)] [Medline: [31501885](https://pubmed.ncbi.nlm.nih.gov/31501885/)]
57. Zaheer M, Guruganesh G, Dubey K, Ainslie J, Alberti C, Ontanon S, et al. Big Bird: transformers for longer sequences. *Proceedings of the 34th Conference on Neural Information Processing Systems (NeurIPS 2020) 2020* [[FREE Full text](#)]
58. Zhang D, Tahami AV, Abualsaud M, Smucker MD. Learning trustworthy web sources to derive correct answers and reduce health misinformation in search. *SIGIR '22: Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval 2022*:2099-2104. [doi: [10.1145/3477495.3531812](https://doi.org/10.1145/3477495.3531812)]
59. The ClueWeb12 Dataset. The Lemur Project. URL: <http://lemurproject.org/clueweb12/> [accessed 2024-04-18]

60. Zuccon G, Palotti J, Goeuriot L, Kelly L, Lupu M, Pecina P, et al. The IR Task at the CLEF eHealth evaluation lab 2016: User-centred health information retrieval. 2016 Presented at: CLEF 2016 - Conference and Labs of the Evaluation Forum; September 5-8, 2016; Évora, Portugal p. 255-266 URL: <https://ceur-ws.org/Vol-1609/16090015.pdf>
61. Bennani-Smires K, Musat C, Hossmann A, Baeriswyl M, Jaggi M. Simple unsupervised keyphrase extraction using sentence embeddings. Proceedings of the 22nd Conference on Computational Natural Language Learning 2018:221-229. [doi: [10.18653/v1/K18-1022](https://doi.org/10.18653/v1/K18-1022)]
62. Ogilvie P, Callan J. Combining document representations for known-item search. SIGIR '03: Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval 2003:143-150. [doi: [10.1145/860462.860463](https://doi.org/10.1145/860462.860463)]
63. Webber W, Moffat A, Zobel J. A similarity measure for indefinite rankings. ACM Trans. Inf. Syst 2010 Nov 23;28(4):1-38. [doi: [10.1145/1852102.1852106](https://doi.org/10.1145/1852102.1852106)]
64. Clarke CLA, Smucker MD, Vtyurina A. Offline evaluation by maximum similarity to an ideal ranking. CIKM '20: Proceedings of the 29th ACM International Conference on Information & Knowledge Management 2020:225-234. [doi: [10.1145/3340531.3411915](https://doi.org/10.1145/3340531.3411915)]
65. Hugging Face. URL: <https://huggingface.co> [accessed 2024-04-18]
66. Zhang B, Naderi N, Mishra R, Teodoro D. Online health search via multi-dimensional information quality assessment based on deep language models. MedRxiv Preprint posted online on January 11, 2024 [FREE Full text] [doi: [10.1101/2023.04.11.22281038](https://doi.org/10.1101/2023.04.11.22281038)]
67. Pradeep R, Ma X, Nogueira R, Lin J. Scientific claim verification with VerT5erini. Proceedings of the 12th International Workshop on Health Text Mining and Information Analysis 2021:94-103 [FREE Full text]
68. Abualsaud M, Chen IX, Ghajar K, Minh LNP, Smucker MD, Tahami AV, et al. UWaterlooMDS at the TREC 2021 Health Misinformation Track. NIST Special Publication 500-335: The Thirtieth Text REtrieval Conference Proceedings (TREC 2021) 2022:1 [FREE Full text]
69. Schlicht I, Paula AD, Rosso P. UPV at TREC Health Misinformation Track 2021 ranking with SBERT and quality. NIST Special Publication 500-335: The Thirtieth Text REtrieval Conference Proceedings (TREC 2021) 2022:1 [FREE Full text]
70. Fernández-Pichel M, Prada-Corral M, Losada DE, Pichel JC, Gamallo P. CiTIUS at the TREC 2021 Health Misinformation Track. NIST Special Publication 500-335: The Thirtieth Text REtrieval Conference Proceedings (TREC 2021) 2022:1 [FREE Full text]
71. Belkin NJ, Kantor P, Fox EA, Shaw JA. Combining the evidence of multiple query representations for information retrieval. Information Processing & Management 1995 May;31(3):431-448. [doi: [10.1016/0306-4573\(94\)00057-A](https://doi.org/10.1016/0306-4573(94)00057-A)]
72. Bondarenko A, Fröbe M, Gohsen M, Günther S, Kiesel J, Schwerter J, et al. Webis at TREC 2021: Deep Learning, Health Misinformation, and Podcasts Tracks. NIST Special Publication 500-335: The Thirtieth Text REtrieval Conference Proceedings (TREC 2021) 2022:1 [FREE Full text]
73. Teodoro D, Mottin L, Gobeill J, Gaudinat A, Vachon T, Ruch P. Improving average ranking precision in user searches for biomedical research datasets. Database (Oxford) 2017 Jan 01;2017:bax083 [FREE Full text] [doi: [10.1093/database/bax083](https://doi.org/10.1093/database/bax083)] [Medline: [29220475](https://pubmed.ncbi.nlm.nih.gov/29220475/)]
74. 2021 Health Misinformation Track. TREC. 2022. URL: <https://trec.nist.gov/data/misinfo2021.html> [accessed 2024-04-18]

Abbreviations

- BERT:** Bidirectional Encoder Representations from Transformers
 - BM25:** Best Match 25
 - C4:** Colossal Clean Crawled Corpus
 - CLEF:** Conference and Labs of the Evaluation Forum
 - CSS:** cascading style sheets
 - nDCG:** normalized discounted cumulative gain
 - NIST:** National Institute of Standards and Technology
 - RBO:** rank-biased overlap
 - RoBERTa:** robustly optimized BERT approach
 - RRF:** reciprocal rank fusion
 - TREC:** Text Retrieval Conference
-

Edited by B Malin; submitted 12.09.22; peer-reviewed by D Carvalho, D He, S Marchesin; comments to author 10.04.23; revised version received 12.07.23; accepted 15.01.24; published 02.05.24.

Please cite as:

Zhang B, Naderi N, Mishra R, Teodoro D

Online Health Search Via Multidimensional Information Quality Assessment Based on Deep Language Models: Algorithm Development and Validation

JMIR AI 2024;3:e42630

URL: <https://ai.jmir.org/2024/1/e42630>

doi: [10.2196/42630](https://doi.org/10.2196/42630)

PMID: [38875551](https://pubmed.ncbi.nlm.nih.gov/38875551/)

©Boya Zhang, Nona Naderi, Rahul Mishra, Douglas Teodoro. Originally published in JMIR AI (<https://ai.jmir.org>), 02.05.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR AI, is properly cited. The complete bibliographic information, a link to the original publication on <https://www.ai.jmir.org/>, as well as this copyright and license information must be included.

Original Paper

Multiscale Bowel Sound Event Spotting in Highly Imbalanced Wearable Monitoring Data: Algorithm Development and Validation Study

Annalisa Baronetto^{1,2}, MSc; Luisa Graf³, MSc; Sarah Fischer^{4,5}, MD, PhD; Markus F Neurath^{4,5}, MD; Oliver Amft^{1,2}, PhD

¹Hahn-Schickard, Freiburg, Germany

²Intelligent Embedded Systems Lab, University of Freiburg, Freiburg, Germany

³Chair of Digital Health, Friedrich-Alexander Universität Erlangen-Nürnberg, Erlangen, Germany

⁴Medical Clinic 1, University Hospital Erlangen, Friedrich-Alexander Universität Erlangen-Nürnberg, Erlangen, Germany

⁵Deutsches Zentrum Immuntherapie, Erlangen, Germany

Corresponding Author:

Annalisa Baronetto, MSc

Hahn-Schickard

Georges-Köhler-Allee 302

Freiburg, 79110

Germany

Phone: 49 761 887 865 ext 732

Email: ab1591@students.uni-freiburg.de

Abstract

Background: Abdominal auscultation (i.e., listening to bowel sounds (BSs)) can be used to analyze digestion. An automated retrieval of BS would be beneficial to assess gastrointestinal disorders noninvasively.

Objective: This study aims to develop a multiscale spotting model to detect BSs in continuous audio data from a wearable monitoring system.

Methods: We designed a spotting model based on the Efficient-U-Net (EffUNet) architecture to analyze 10-second audio segments at a time and spot BSs with a temporal resolution of 25 ms. Evaluation data were collected across different digestive phases from 18 healthy participants and 9 patients with inflammatory bowel disease (IBD). Audio data were recorded in a daytime setting with a smart T-Shirt that embeds digital microphones. The data set was annotated by independent raters with substantial agreement (Cohen κ between 0.70 and 0.75), resulting in 136 hours of labeled data. In total, 11,482 BSs were analyzed, with a BS duration ranging between 18 ms and 6.3 seconds. The share of BSs in the data set (BS ratio) was 0.0089. We analyzed the performance depending on noise level, BS duration, and BS event rate. We also report spotting timing errors.

Results: Leave-one-participant-out cross-validation of BS event spotting yielded a median F_1 -score of 0.73 for both healthy volunteers and patients with IBD. EffUNet detected BSs under different noise conditions with 0.73 recall and 0.72 precision. In particular, for a signal-to-noise ratio over 4 dB, more than 83% of BSs were recognized, with precision of 0.77 or more. EffUNet recall dropped below 0.60 for BS duration of 1.5 seconds or less. At a BS ratio greater than 0.05, the precision of our model was over 0.83. For both healthy participants and patients with IBD, insertion and deletion timing errors were the largest, with a total of 15.54 minutes of insertion errors and 13.08 minutes of deletion errors over the total audio data set. On our data set, EffUNet outperformed existing BS spotting models that provide similar temporal resolution.

Conclusions: The EffUNet spotter is robust against background noise and can retrieve BSs with varying duration. EffUNet outperforms previous BS detection approaches in unmodified audio data, containing highly sparse BS events.

(JMIR AI 2024;3:e51118) doi:[10.2196/51118](https://doi.org/10.2196/51118)

KEYWORDS

bowel sound; deep learning; event spotting; wearable sensors

Introduction

There are various diagnostic tools to assess bowel motility, including questionnaires, ultrasound, and endoscopic examinations [1,2]. However, there is a lack of computational tools to monitor digestion continuously across the gastrointestinal tract. Abdominal examinations using a stethoscope (ie, auscultation of the bowel) is a common clinical practice to interpret bowel sounds (BSs) [3]. While BSs could help examiners perform diagnoses [4,5], auscultation is mostly done for a few minutes only [6]. However, BSs occur sparsely over time, have varying patterns, and often exhibit low volume. Previous investigations (eg, [7]) recommended recording BSs with multiple sensors and over longer periods to maximize the amount of BS observations. Craine et al [8] reported that changes in BS occurrences across different digestive phases were statistically different in patients with irritable bowel syndrome and Crohn disease (CD). Later studies (eg, [9]) showed that digestion analysis based on BSs could support bowel motility assessment as well as monitoring food intake. For instance, an increased number of BS events could indicate bowel hyperactivity, caused by, for example, gastroenteritis or inflammatory bowel disease (IBD) [6]. Yao and Tai [10] recorded BSs across patients with CD, patients with ulcerative colitis (UC), and healthy controls. The authors reported that patients with CD showed the highest BS peak frequency, while patients with UC had the highest BS event count per unit time. Consequently, spotting BS occurrences in continuous audio could provide important information to assess digestion. To date, however, the clinical assessment based on BS remains qualitative and lacks quantification of BS characteristics [11]. For all of the aforementioned applications, short manual auscultation is considered challenging, as it provides examiners with insufficient information on dynamic bowel conditions.

Various wearable prototypes were proposed to record BSs in healthy volunteers and patients with digestive disorders (eg, [12,13]). Study protocols were primarily designed to observe BSs under controlled laboratory settings, that is, while participants laid down and rested, to minimize noise artifacts. The large amount of audio data that could be recorded by wearable systems renders a manual analysis infeasible.

Previous studies (eg, [14]) have attempted to reduce the amount of audio data to be manually analyzed with segment-based approaches that detected audio sections containing BS events. Moreover, methods were proposed to improve BS event detection and ease expert examination, by determining the onset and offset of the BS patterns in audio data streams (eg, [9,15]). Nevertheless, most algorithms were tested on balanced data sets or selected subsets of the recordings only, from dozens of minutes to a few hours. However, when collecting data with a wearable device, the BS ratio of relevant events, for example, BSs versus other surrounding sounds, largely influences retrieval performance, which reflects a basic problem in pattern spotting [16]. Specifically, in naturalistic, unmodified audio data, BS events appear sparsely and their low amplitude compared with other body sounds, for example, lung sounds, hampers BS spotting. For example, Ficek et al [15] reported that temporal

sparsity of BSs could increase the false-positive rate. Previous studies have shown that BSs can vary in duration, from dozens of milliseconds to a few seconds [17,18]. Hence, the key challenge is to spot BS events, embedded in a large amount of irrelevant audio data, commonly referred to as the NULL class. To spot very short BS events (ie, those <100 ms), detection algorithms need to maximize temporal resolution, which is usually done by minimizing the sliding window size used to inspect the data stream. However, reducing the sliding window size removes context from the audio data, and thus may not improve recognition performance for BSs far shorter than 1 second.

In this paper, we present a BS spotting method based on a deep neural network (DNN) model. Our DNN model spots BS events by analyzing a continuous data stream recorded with a wearable device at 10-second audio segments. Using a multiscale approach, we can retrieve BS event onset and offset at a temporal resolution, that is, the smallest prediction duration, of 25 ms. Our approach is inspired by the way humans perform auscultation: particularly, for BS shorter than 1 second, experts would listen to the audio data surrounding the BS event to obtain an acoustic context. We evaluated our spotting approach on continuous BS recordings collected across different digestive phases, including sedentary activities and food intake. Unlike previous studies, we tested our approach on audio data having natural BS temporal distribution, that is, no resampling was applied to our data set.

The paper provides the following contributions:

- We present a DNN-based method for BS spotting in continuous data streams. Our model achieves a temporal resolution of 25 ms through a multiscale approach.
- We evaluate our model on 136 hours of annotated audio data recorded from 18 healthy participants and 9 patients with an IBD, in total including more than 11,000 annotated BS events. To spot BS events, we do not discern between healthy controls and patients with IBD, but focus on common BS acoustic properties across the different bowel conditions.
- We analyze spotting errors over the unmodified audio data streams. In addition, we analyze our model's performance under various signal-to-noise ratios (SNRs), for different BS event durations, and by varying the temporal sparsity of BS events (ie, BS ratio).

Methods

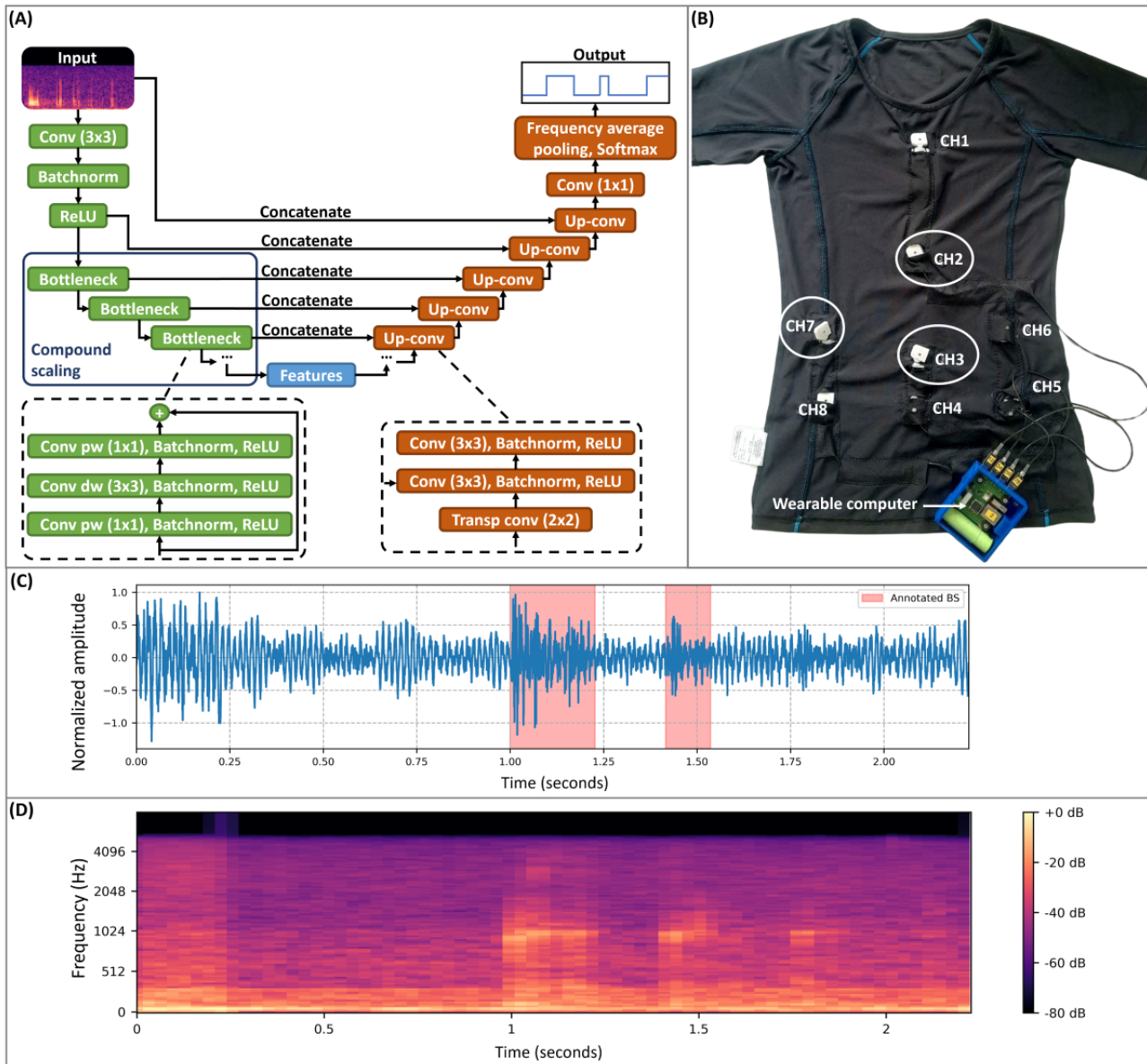
Overview

Here, we describe the DNN model proposed for BS detection. Subsequently, we detail the spotting procedure, the BS evaluation study, and our evaluation methods.

Efficient-U-Net Model

Figure 1 illustrates the Efficient-U-Net (EffUNet) DNN model architecture. The proposed model was based on UNet [19] and EfficientNet [20] models, hence the name EffUNet. In total, EffUNet has approximately 18.1 million parameters.

Figure 1. (A) Architecture of the proposed Efficient-U-Net (EffUNet) for bowel sound (BS) spotting. The model took an audio spectrogram as an input and extracted relevant features with EfficientNet-B2 during the encoding (green boxes). Subsequently, features were decoded, that is, upsampled and concatenated with higher-resolution features to locate them on the original spectrogram (orange boxes). Finally, the obtained 2D features were converted to a BS detection mask by applying average pooling along the frequency dimension and a Softmax operation to the obtained 1D temporal maps. Spectrogram frames with highest BS class probability were identified as containing BS. (B) Inside out of the smart T-shirt showing the embedded electronics. Microphones and the wearable computer were protected and isolated by 3D printed covers. Microphones CH2, CH3, and CH7 (white circles) were used during BS annotation. (C and D) Illustration and time-frequency representation of 2 expert-annotated BS events in the continuous data stream collected from 1 study participant with very different BS event duration. batchnorm: batch normalization; Conv: convolution; dw: depthwise; pw: pointwise; ReLU: rectified linear unit; transp: transposed; up-conv: transposed convolution.



UNet is a convolutional neural network (CNN) that was originally proposed for biomedical image segmentation [19]. The model name is given by its U-shaped architecture, which is composed of an encoder followed by a decoder network. The encoder extracts relevant features from the DNN input, and the decoder generates a segmentation mask by upsampling features from the encoder's last layer and concatenating them with higher-resolution features extracted from the encoder's earlier layers. Each block of the decoder is therefore composed of a 2×2 transposed convolution (up-conv), followed by two 3×3 convolutions. Upsampling restores the original input resolution, and the concatenation improves the localization of the extracted features. A final convolutional layer classifies each input point,

for example, each time-frequency bin of audio spectrogram F_k , with \boxed{x} . We based our approach on the UNet architecture because of its high classification resolution compared with the input data size (ie, it could classify images by pixels). Other common CNN architectures (eg, [21]) usually provide 1 prediction per input data, for example, predict object presence in an image, thus omitting other relevant information, such as the object location in the image. In audio processing, architectures similar to UNet have been used mainly for source-separation tasks [22].

In computer vision segmentation tasks, the model output is usually a 2D map with the same dimensions as the input data.

In our work, EffUNet takes an audio spectrogram as an input and returns a binary detection mask in the time domain. To obtain a 1D mask, we applied an average pooling along the frequency dimension and a Softmax function to the model output. EffUNet thus classified each spectrogram time bin F_k as containing either BS or non-BSs (NBSs). Thus, the spotting temporal resolution corresponds to the audio spectrogram frame length.

As an encoder, several CNN architectures were used and tested in computer vision to improve model performance (eg, residual network [21]). In our work, we used EfficientNet [20] as an encoding model. Similar architectures based on combined EfficientNet and UNet were already proposed in computer vision tasks with promising results (eg, [23]). EfficientNet architectures were introduced to improve image classification performance while reducing the amount of model parameters. The simpler architecture compared with other common CNNs makes EfficientNet suitable for mobile and edge computing applications. In EfficientNet, convolution operations are performed by a bottleneck block: (1) an inverted bottleneck (1×1) convolution, (2) a depthwise (3×3) convolution to extract features, and (3) a pointwise (1×1) convolution to linearly combine the features. Similarly, to standard convolutional layers, a batch normalization layer and a linear layer with rectified linear unit activation are applied after each convolution. In addition, residual connections are added between bottleneck blocks. Different EfficientNet configurations are available with compound scaling, that is, simultaneous increase of features count, number of layers, and input data resolution. We selected the EfficientNet-B2 configuration for our detection model. The EfficientNet-B2 architecture was already used in audio tagging tasks with promising performance [24].

EffUNet Spotting Procedure

Data Preprocessing and Training Pipeline

From audio data, log-Mel spectrograms were extracted to train and evaluate EffUNet. Here, we detail the data preprocessing and training pipeline, including transfer learning and data augmentation. Subsequently, we describe the spotting implementation.

Audio Preprocessing

Recordings were filtered with a high-pass biquadratic filter (cutoff: 60 Hz) to remove signal offset. Subsequently, recordings were split into nonoverlapping audio segments S_i with duration $\delta=10$ seconds. Each audio channel was preprocessed for BS spotting independently. We defined each audio segment S_i as a set of samples:



where \square_x is a time series sample.

Each audio segment was converted to a log-Mel spectrogram using 128 frequency bins, a sliding 25-ms window, and a stride length of 10 ms. As described in the "Efficient-U-Net Model" section, the 25-ms window corresponds to the spotting temporal resolution of EffUNet. Hanning windowing was applied to the

sliding windows. Each resulting spectrogram had 128 Mel bins and 998 frames. According to EfficientNet-B2 pretraining [24], we zero-padded the spectrogram along the time axis to obtain 1056 time bins. The obtained spectrograms were standardized.

For every S_i , we defined the audio spectrogram time bins F_k as follows:



where \square_x is a time series sample. The sliding window duration $\gamma=25 \text{ ms} \cdot f_s$ and the stride length $\sigma=10 \text{ ms} \cdot f_s$ in all time series samples.

Every annotated BS event e_j can be denoted as a set of time series samples as follows:



where $t_{j,s}$ and $t_{j,e}$ are the onset and offset of BS event e_j in time series samples, respectively.

For model learning, BS manual annotations were converted to audio spectrogram ground truth masks according to the approach outlined by Ficek et al [15]: a spectrogram frame F_k was defined as containing BSs ($F_{k,BS}$) if the time overlaps \square_x between the spectrogram frame and a BS event e_j was $\geq 50\%$. Thus, for set $F_{k,BS}$:



where $\square_x=0.5$ is the temporal overlap and $|\cdot|$ is the set cardinality. Otherwise, the spectrogram frame was denoted as containing NBSs ($F_{k,NBS}$), that is, the NULL class. We define supersets of all spectrogram frames as $F_{k,BS} \square_x F_{BS}$ and $F_{k,NBS} \square_x F_{NBS}$. Therefore, for each audio segment S_i , we obtained from EffUNet a binary mask M_i denoted as F_k . As with the log-Mel spectrograms, we zero-padded the binary masks M_i along the time axis to obtain 1×1056 time bin masks.

Transfer Learning

The EffUNet encoder, that is, EfficientNet-B2, was initialized with pretraining weights from AudioSet [25]. AudioSet is to date the largest audio data set, containing over 500 audio classes with over 2 million 10-s audio clips (ie, >5000 hours of audio data). AudioSet contains sound examples from daily living, including, among others, speech, environmental sounds, and BSs. We believe that pretraining on a large variety of sounds could improve the spotting robustness against background noise and other artifacts. As both AudioSet and the BS recordings of our study were sampled at the same frequency (ie, 16 kHz), the pretrained encoder feature extraction was compatible with our BS data. Nevertheless, because no onset and offset of audio events were originally provided in AudioSet, no pretraining could be applied to our EffUNet decoder. Therefore, the decoder weights were initialized with He initialization [26].

Training and Data Augmentation

EffUNet training parameters were selected according to the Pretraining, Sampling, Labeling, and Aggregation pipeline [24]. After model initialization, EffUNet was trained for 25 epochs using an imbalanced batch size of 32 and an initial learning rate of 1×10^{-4} . The learning rate was subsequently reduced with a decay of 0.85 for each epoch, starting from the sixth epoch. Adam optimizer [27] was used with weight decay of 5×10^{-7} , $\beta_1=0.95$, $\beta_2=0.999$. We used the following loss function for optimization:

$$\mathcal{L} = \mathcal{L}_{CE} + \mathcal{L}_{Dice}$$

where \mathcal{L}_{CE} is the cross-entropy loss [28] calculated between the prediction \hat{y} and the ground truth y , and \mathcal{L}_{Dice} is the dice loss [29]. While the cross-entropy loss maximizes the model performance to classify single spectrogram time bins, the dice loss maximizes the similarity between the predicted binary mask and the ground truth (ie, expert BS annotation).

During the training, the input audio spectrograms were randomly transformed to improve the model generalization. On each batch, time-frequency masking [30] was applied to up to 24 frequency bins and up to 10% of the time bins. In addition, spectrograms were randomly shifted along the time axis with a maximum shift of +10 or -10 time bins. Random white noise with

magnitude in the range [0, 0.1) was also added to the input spectrogram. For the evaluation, we selected the model weights obtained at the end of the training.

Spotting Implementation

The binary masks M_i obtained from EffUNet were converted to onset/offset predictions of BS events. Spotted BS events $\mathcal{D}_{i,BS}$ were described as follows:

$$\mathcal{D}_{i,BS} = \{t \in \mathcal{T} \mid M_i(t) = 1\}$$

$$\mathcal{D}_{i,BS} = \{t \in \mathcal{T} \mid M_i(t) = 1\}$$

where \mathcal{T} is a time series sample, and $D_{i,BS}$ is the set of N consecutive overlapping audio spectrogram time bins \mathcal{T} that were detected as containing BSs.

Evaluation Study and Data Preprocessing

Study Protocol

The study involved 27 participants (13 females, aged 21-69 years; clothing sizes: S-XL; and BMI 17.2-32.2 kg/m²). Among the 27 individuals, 9 were patients with IBD. Table 1 illustrates the population characteristics of our data set. After signing written consent, participants were invited to the laboratory in the morning before breakfast.

Table 1. Characteristics of the population included in this study.

Cohort	IBD ^a	UC ^b	CD ^c	IBD activity	IBD remission	Healthy	Total
Participants, n	9	6	3	6	3	18	27
Sex, n							
Male	3	1	2	5	2	11	14
Female	6	5	1	1	1	7	13
Age (years), median (range)	39 (23-69)	36 (23-69)	39 (39-47)	33 (23-69)	47 (45-58)	28 (21-56)	28 (21-69)
BMI (kg/m ²), median (range)	24.6 (17.9-26.2)	22.8 (17.9-26.0)	24.6 (24.2-26.1)	24.4 (17.9-25.3)	26.0 (18.4-26.1)	22.3 (17.1-32.2)	22.5 (17.2-32.2)

^aIBD: inflammatory bowel disease.

^bUC: ulcerative colitis.

^cCD: Crohn disease.

A smart T-shirt (GastroDigitalShirt) [31] was used to record BSs from 8 embedded digital miniature microphones (SPH0645LM4H-B; Knowles) aligned on the abdomen. Microphones were positioned according to the 9-quadrant reference abdominal map and arranged to follow the digestive process. For example, the first channel was placed on the esophagus, the second channel on the stomach. A belt-worn computer collected and saved all microphone channels at $f_s=16$ kHz. A tight-fitting design and various sizes were used to ensure comfort and optimal skin attachment. The fabric was based on elastane, thus highly stretchable. The cloth cut was based on a compression T-shirt to minimize noise artifacts as a result of motion. Different cloth cuts for females and males were prepared to fit all body shapes and provide optimal comfort. Figure 1 shows our wearable prototype and the embedded electronics.

Participants were asked to put on the smart T-shirt and audio was continuously recorded from 1 hour before breakfast (fasting phase) to 1 hour after breakfast (postprandial phase). To avoid abnormal bowel motility stimulation, induced by, for example, physical movements [32], participants laid down and quietly relaxed when there was no meal intake or other activity. They were recommended to read a book, watch or listen to multimedia on a tablet, or sleep. Although participants were relaxing, they could move on the lounge chair, if desired. Moreover, participants were required to stand up and sit down as per the study protocol, so motion artifacts could be included in the recording. While eating breakfast, participants sat at a table and were allowed to talk to the study personnel or move freely around the room. The audio was continuously recorded during the whole session. Participants were allowed to drink water throughout the recording and could pause it anytime for a break

(eg, to visit the toilet). Along with BSs, other sound events could be captured (ie, NULL class data). For instance, conversations between participants and study personnel and other environmental sounds from the room surroundings, for example, traffic as well as activities and voices outside the recording room, could be recorded. In addition, voluntary body position adjustments and eating or drinking could introduce noise artifacts.

Upon completing the recording protocol, study participants were asked to rate the T-shirt's comfort and usability to confirm that it could be worn for the recording duration. The assessment was based on the wearable comfort assessment questionnaire [33]. The study participants reported no discomfort caused by the embedded electronics.

BS Annotation

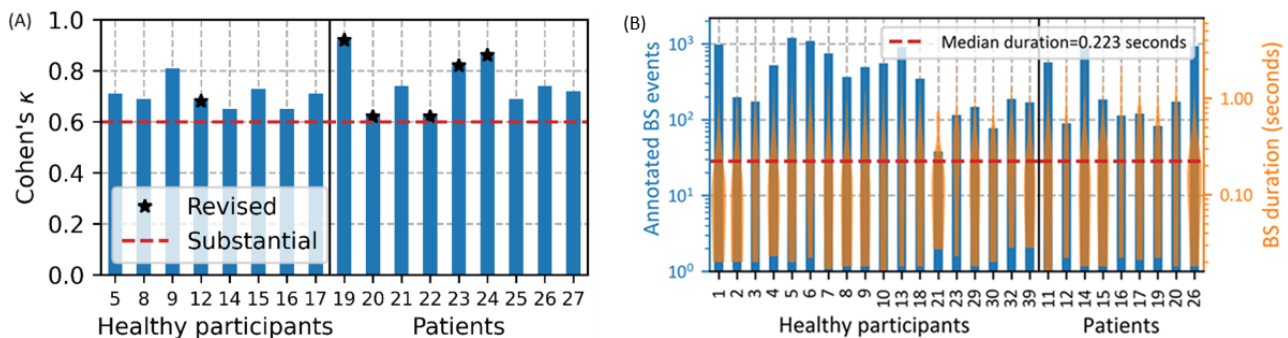
Recordings were annotated by pairs of raters through audio and visual inspection of the raw audio data using Audacity (The Audacity Team). Annotations were sample specific, that is, no quantization of the BS events' onset/offset was performed. An example of annotated BS events with different durations is presented in Figure 1. For raters to identify BS events with varying durations and amplitudes in the recordings, view time resolution in Audacity had to be adjusted. On average, each rater required 8-12 hours to label 1 hour of recording, depending on event rate (ie, the number of BSs per unit time) and noise level. As a result of the time-consuming annotation process, only a subset of the recordings was annotated by more than 1 rater to evaluate interrater agreement (see below). The remaining data were labeled by 1 of the raters and the annotations were checked by the other rater. Among all participants, audio data from the sensors positioned on the stomach (CH2) and the small intestine above the navel (CH3) were annotated (see Figure 1 for a sensor map). As IBD usually affects the distal part of the gastrointestinal tract, in the patient group and in some

individuals from the healthy group, an additional microphone placed on the distal part of the large intestine (CH7) was included in the annotation to evaluate our spotting approach with additional BS patterns. Because of SNR limitations, the channel located at the large intestine could not be annotated for all participants. The annotation was performed on each recording channel separately because BSs could be recorded at 1 or more locations depending on the sound propagation across the abdomen. The position of annotated audio channels on the T-shirt is shown in Figure 1.

Based on BS features reported in the literature [17,18], as well as preliminary auscultation sessions, and early annotation reviews, labeling guidelines were selected and agreed upon between raters: BS duration had to be 18 ms or more, and consecutive BS events with sound-to-sound interval less than 100 ms were marked as a single event. Noisy or BSs not visible in the audio signal were labeled as tentative.

Cohen κ interrater agreement was used to evaluate the annotation quality. Two raters labeled the first 30 minutes of recordings from 8 healthy participants and 9 patients. After the agreement evaluation for each participant's recording, a label review session was conducted to discuss and revise any BSs with disagreement. If an agreement of $\kappa < 0.6$, indicating slight to moderate disagreement, was observed for a participant data set, then the agreement score was recalculated based on an additional 10 minutes of the participant's recording after the review and revision. Overall, in the healthy group and the patient group, agreement on nontentative BS annotations was substantial, with Cohen κ of 0.70 and 0.75, respectively. As the data imbalance between BSs and NBSs increases, the maximum achievable agreement between raters decreases. Therefore, agreements are deemed fair to good beyond chance for scores between 0.40 and 0.75 [34]. Figure 2 illustrates the interrater agreement per study group.

Figure 2. (A) Interrater agreement per study group. The evaluation was performed on a subset of the study participants. Overall, the agreement on the nontentative bowel sound (BS) annotations was substantial. (B) Amount and duration distribution of BS annotated per participant. Most BSs are short (median duration 223 ms).



Overall, 11,482 BSs plus 3801 tentative BSs were annotated on approximately 136 hours of audio. The annotated BSs had a total duration of 1.22 hours, with 52.39 minutes recorded from the healthy group and 20.71 minutes recorded from the patient group. Of the total nontentative annotated BSs, 3215 were observed at the stomach, 5667 at the small intestine, and 2600 at the large intestine. Because of the noisy signal patterns that could affect the spotting performance, tentative BSs were not included in the analysis and, therefore, were considered NBSs

(ie, belonging to the NULL class). The quantity and duration of annotated BSs across all participants are shown in Figure 2. As reported by previous studies [17,18], BS event duration ranges from 18 ms to a few seconds. However, most annotated BSs have a very short duration (<500 ms).

Evaluation Methods

Validation Method

Leave-one-participant-out (LOPO) cross-validation (CV) was used to evaluate spotting performance: audio data from all but 1 participant were used to train the DNN, and its performance was evaluated on the excluded participant's data. Performance statistics were obtained from the results of each validation set.


Evaluation Metrics



Precision and recall (PR) metrics and F_1 -score were used to evaluate spotting performance across all testing data. Metrics were calculated using a samplewise approach based on Mesaros et al [35], that is, model predictions and BS annotations were compared sample-by-sample: $t_{i,S}=1$ ($f_S \approx 0.06$ ms). Thus, our evaluation approach was independent of the spotting algorithm resolution. Furthermore, we compared directly with BS annotations without considering their spotting frame-adjusted versions.

We analyzed spotting detection errors with the 2-class segment error metric [36]. False-positive FP_i were marked as merge errors if they connected 2 consecutive events e_j , overflow errors if FP_i occurred at the beginning or end of an event e_j , and insertion errors otherwise. False-negative FN_i were marked as fragmentation errors if FN_i occurred within 1 event e_j , underfill errors if they occurred at the beginning or end of an event e_j , and deletion errors otherwise. For each LOPO fold, we derived the overall detection timing errors as time duration.

Model performance statistics were described using median and IQR values. IQR was determined as the difference between quartile Q1 (ie, the mid value between the median and the minimum) and quartile Q3 (ie, the mid value between the maximum and the median). We further evaluated spotting performance by analyzing PR metrics over SNR as follows:



where θ_S is an SNR threshold applied to audio segments S_i . For each S_i , SNR was computed in the log-decibel scale as the ratio between the signal power of e_j  S_i and the background noise in S_i .

Moreover, we analyzed PR metrics over BS duration. To estimate TP_i , FP_i , and FN_i depending on BS duration, we only considered annotated events e_j and detected event  so that $|e_j| \geq \theta_D$ and , where θ_D was a BS duration threshold.

Retrieval generalization was additionally evaluated by analyzing PR over event rate. Event rate was defined as BS events per time unit according to Amft [16]. To compare our model performance with related work, we converted event rate to BS ratio (ie, the ratio between spectrogram time bins containing BSs and total time bins) as follows:



For each validation fold, we swept the BS ratio from 0.00001 to 0.60 by randomly sampling K from $F_{k,BS}$ and J from $F_{k,NBS}$ so that BS ratio = $K/(K+J)$, thus corresponding to bootstrap samples according to the count of validation folds. For each selected BS ratio, we calculated the corresponding event rate per hour of recording. Although models were not retrained on the selected BS ratios, the analysis provides insights into the performance at different class imbalance levels. We show that spotting performance depends on the BS ratio and compared our results with published works in the literature, which mostly focused on artificially balanced data sets.

Comparison With Prior Work

We examined the performance of existing models for BS detection on our data set. For comparison purposes, we focused on spotting models with similar temporal resolution. Segment-based spotting approaches, such as those described by [37], were excluded from our comparison because of their distinct design scope, which does not include providing BS event onset/offset. Among recent works, the convolutional recurrent neural network (CRNN) by Ficek et al [15], the CNN by Wang et al [9], and the CNN by Kutsumi et al [38] offer the highest temporal resolution. The CRNN training pipeline, originally evaluated in a data set of 53 minutes, could not be scaled to our substantially larger data set, as the CRNN optimization did not converge on our highly imbalanced data set. The CNN by Kutsumi et al [38] could not be reimplemented as it lacked methodological information (see the "Discussion" section). We, therefore, reimplemented and trained the CNN by Wang et al [9] for 30 epochs using an initial learning rate of 0.001 and a balanced batch size of 128. Adadelta optimizer with a weight decay of 10^{-7} and a decay rate of 0.95 was used to optimize the cross-entropy loss function. Unlike our EffUNet model, the CNN takes as input a log-Mel spectrogram extracted from 60 ms nonoverlapping audio segments using a 50-ms sliding window (preprocessed with Hanning windowing) and a stride length (σ) of 5 ms. The CNN classified each spectrogram as either containing BSs or not. We split our data accordingly and assigned each 60-ms audio segment to either the BS or the NBS class using Equation 4. We could not follow the labeling approach proposed by Wang et al [9] because the authors manually annotated each 60-ms audio segment individually rather than the continuous data stream. The acquired audio data were preprocessed by applying a high-pass filter with a cutoff of 80 Hz. According to the authors, spectrograms were standardized and no data augmentation was used during the training. To directly compare the results with EffUNet, the CNN was evaluated using LOPO CV.

Ethics Approval

The study was approved by the Ethics Commission of the Friedrich-Alexander Universität Erlangen-Nürnberg (protocol number 73_20 B).

Results

BS Ratio and F_1 -Scores

Based on the annotated 11,482 BS events, we obtained a BS ratio of approximately 0.0089 for the data set. Figure 3 shows

F_1 -scores across all participants and for each study group. EffUNet achieved the largest F_1 -score in the healthy group. For both groups, a median F_1 -score of 0.73 was obtained. Although the patient group yields the lowest IQR, an outlier with an F_1 -score of approximately 0.50 was identified.

Figure 3. (A) Box plots of F_1 -scores across all participants and study groups. For both groups, a median F_1 -score of 0.73 was obtained; however, the patient group showed the lowest IQR. (B) Precision and recall (PR) over signal-to-noise ratio (SNR) analysis. The number of bowel sound (BS) events considered for each threshold is also shown. When the SNR is >4 dB, more than 80% of BSs were detected by Efficient-U-Net, with a precision in the range of 77%-86%. (C) PR over BS duration analysis. The number of BS events considered for each threshold is also shown. Even when including very short BSs in the analysis, our model could detect events with nearly 75% PR. (D) PR over BS ratio. Dots and error bars show median and IQR, respectively. In our data set, the BS ratio is only 0.0089. Nevertheless, 73% of BSs were recognized with 72% precision. Most studies in the literature were performed on a balanced data set. If the BS ratio was >0.05 , our model would detect BSs with precision $>83\%$.

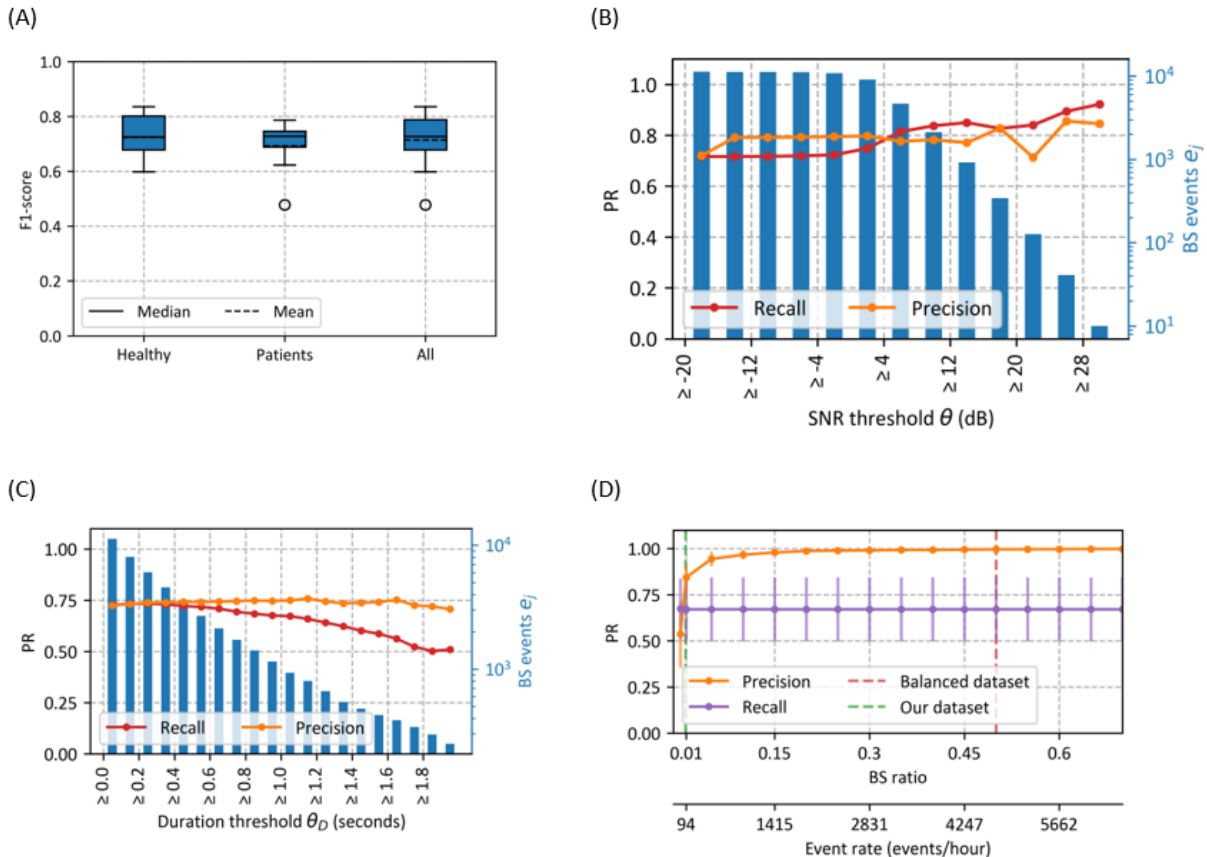


Table 2 shows the median PR of our spotting model for all study groups. The BS spotting achieved identical median precision scores for both the healthy and patient groups. However, BSs

recorded from patients proved more challenging to detect, resulting in a lower median recall compared with the healthy group.

Table 2. Spotting performance for all study groups. While Efficient-U-Net shows the same median precision for both study groups, the median recall was higher for healthy individuals than for patients.

Study group	Precision, median (IQR)	Recall, median (IQR)	F_1 -score, median (IQR)
Healthy	0.80 (0.19)	0.75 (0.19)	0.73 (0.13)
Patients	0.80 (0.23)	0.66 (0.14)	0.73 (0.09)
All	0.80 (0.19)	0.73 (0.18)	0.73 (0.11)

PR metrics and F_1 -score per participant and BS ratio are shown in Figure 4. Overall, BSs were sparser in the patient group than in the healthy group, with a peak BS ratio of 0.015 versus 0.032. Regardless of the BS temporal distribution, EffUNet achieved an F_1 -score above 60%, except for an outlier in the patient group. In participants, where F_1 -score dropped to approximately

60%, the performance decrease was mainly a result of a drop in precision, while most BSs could be retrieved. Figure 4 also shows PR metrics per sensor location. For all locations, EffUNet yielded comparable median and IQR for precision. The median recall was similar for all sensor locations, while recall IQR was largest on the large intestine.

Timing errors using the 2-class segment error analysis across the study groups are shown in Table 3. For both healthy participants and patients, insertion and deletion timing errors were the largest, whereas fragmentation and merge errors were

the lowest. Besides per-participant timing error medians and IQR, the timing error totals are shown. The total errors over the 136 hours of data were 15.54 minutes for insertions and 13.08 minutes for deletions.

Figure 4. (A) Precision and recall (PR) metrics and F_1 -score per study participant. Bowel sound (BS) ratios per participant recording are indicated. Participants whose sensor on the large intestine was annotated are marked by an asterisk. Overall, BSs were sparser in the patient group than in the healthy group, with a peak BS ratio of 0.015 versus 0.032. The F_1 -score was above 60%, except in individuals in whom performance decreased due to a precision drop. (B) PR metrics comparison across the different sensor positions. For all locations, Efficient-U-Net yielded comparable median and IQR for precision. The median recall was similar for all sensor locations, while the recall IQR was largest on the large intestine.

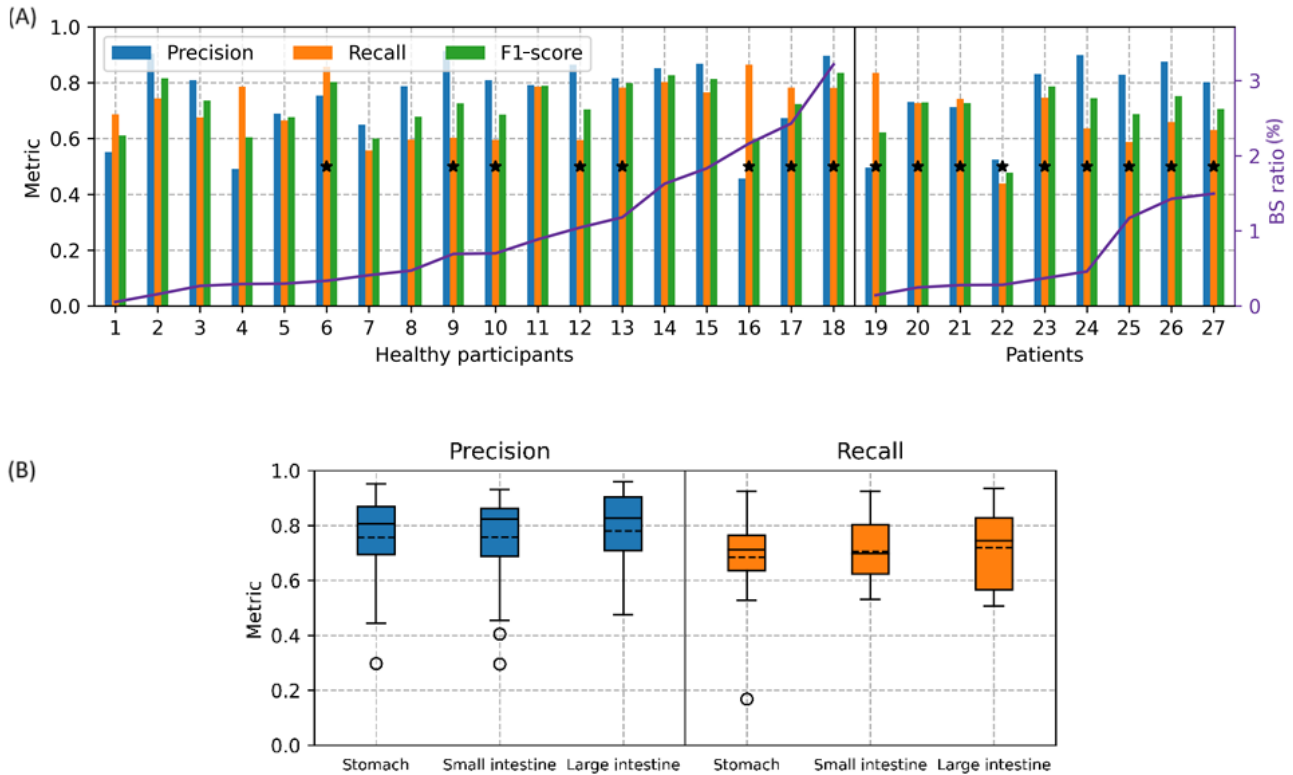


Table 3. Spotting timing errors per participant and totals using 2-class segment error analysis. Overall, insertion and deletion errors showed the largest timing deviations for both healthy and patient groups, whereas fragmentation and merge errors showed the smallest deviations. On our data set of approximately 84 hours for the healthy group and 52 hours for the patient group, 52.4 and 20.7 minutes were annotated as bowel sounds, respectively.

Study group	Insertion	Deletion	Fragmentation	Merge	Overfill	Underfill
Per-participant summed spotting errors (seconds), median (IQR)						
Healthy	15.2 (13.0)	24.0 (35.2)	1.1 (3.4)	0.5 (1.4)	4.2 (9.4)	7.2 (14.8)
Patients	13.1 (14.1)	20.8 (57.5)	1.1 (3.3)	0.7 (0.8)	3.6 (7.4)	7.4 (21.6)
All	14.6 (13.0)	21.9 (36.5)	1.1 (3.4)	0.5 (1.0)	4.1 (7.0)	7.4 (15.4)
Total per-participant summed spotting errors (minutes), median						
Healthy	13.3	7.9	0.8	0.4	3.1	3.7
Patients	2.3	5.2	0.3	0.1	0.9	1.8
All	15.5	13.1	1.1	0.5	4.0	5.6

Figure 3 shows PR metrics over SNR. Annotated BS events considered within each SNR threshold θ_s bin are indicated. Our model detected BSs under different noise conditions with 0.73 recall and 0.72 precision. When BSs are louder than background noise, that is, $\text{SNR} > 4$ dB, more than 83% of BSs were recognized, with precision in the range between 77% and 86%.

PR metrics over BS duration are shown in Figure 3. Even when including very short BSs in the analysis, our EffUNet model

could detect events with nearly 75% recall and precision. EffUNet recall dropped below 60% for BS duration of 1.5 seconds or more, probably because of fragmented predictions that were removed by the duration analysis procedure, that is, BS event duration below θ_D (see the ‘‘Evaluation Metrics’’ section for details).

Figure 3 shows PR metrics over different BS ratios. At our data set’s BS ratio of 0.0089, 73% of the BSs were recognized with

72% precision. Most studies in the literature were performed on a balanced data set. If the BS ratio of our data set was over 0.05, our model would detect BSs with precision greater than 83%.

Comparison With Prior Work

Table 4 shows a comparison of the proposed model with other methods from related work. Our approach was developed and tested on a large data set of 136 hours of recordings. Unlike other studies, the model was tested on the full, highly imbalanced data set. Despite a window of 10 seconds being fed

to EffUNet during the detection, our multiscale approach can detect BSs with a temporal resolution of 25 ms. The CNN proposed by Wang et al [9] yielded a recall of 90% when tested on a balanced data set of approximately 11 minutes in total. On our substantially larger and highly imbalanced data set, however, the CNN model of Wang et al [9] only yielded a median precision of 5% (IQR 0.07) and a median recall of 74% (IQR 0.07). Optimization of the model proposed by Wang et al [9] to deal with imbalanced data may be feasible, but is beyond the scope of this work.

Table 4. Comparison of our model performance with other methods proposed in the literature. Stated performances were those provided by the corresponding articles.

Model	Evaluation data set size	Bowel sound ratio	Recording conditions	Temporal resolution	Precision	Recall
CRNN ^a [15]	N/A ^b	0.0246	Nocturnal recording, clinic	10 ms	0.58	0.86
CRNN [15]	≈11 minutes	0.15	Nocturnal recording, clinic	10 ms	0.83	0.77
CNN ^c [9]	≈15 minutes	0.50	Quiet room	60 ms	N/A	0.90
CNN [38]	2.4 hours	N/A	N/A	100 ms	0.71	0.75
LSTM ^d [39]	≈5 hours	0.45	House rooms	1 second	≈0.94	≈0.99
Autoencoder [37]	≈81 minutes	0.50	Anechoic chamber, synthetic noise	5 seconds	0.92	0.50
Ensemble CNN [40]	49 minutes	0.50	Neonatal intensive care unit	6 seconds	0.97	0.98
CNN + Attention [14]	84 hours	0.15	Laboratory room	10 seconds	0.81	0.70
Efficient-U-Net (this work)	≈136 hours	0.0089	Laboratory/clinical room	25 ms	0.80	0.73

^aCRNN: convolutional recurrent neural network.

^bN/A: not applicable.

^cCNN: convolutional neural network.

^dLSTM: long short-term memory neural network.

Discussion

Principal Findings

Acoustic abdominal monitoring requires physicians to analyze BSs across different digestive phases to detect gastrointestinal disorders. Our data set comprises approximately 2 hours of continuous audio data for each of the 27 participants. We recorded various phases of digestion, from the fasting stage to the food ingestion and consequent postprandial phase. To evaluate the potential of our model in a realistic scenario, the BS natural temporal distribution was left unaltered, that is, no class resampling was applied to the data set. In addition, various activities that are typical of free living were recorded in the study, for example, eating or transition movements (ie, getting up/laying down). While participants laid in a relaxed position for part of the recording session to minimize motion-induced peristalsis stimulation [32], their actions were not constrained (eg, they could grab a bottle and drink water if desired). In particular, participants were allowed to freely move and talk during breakfast. Moreover, as the recording room was not acoustically isolated from the surroundings, various noise sources could be captured in the recording besides the artifacts introduced by the participant movements (see the “Study Protocol” section). We believe that our recording setting and

data amount can be considered as a realistic representation of common activities and sedentary lifestyles. While we suggested a sedentary behavior for participants to obtain nonstimulated BS distributions (as discussed earlier), future work could evaluate BS spotting under different conditions, such as specific physical activities, sports, and stress. If necessary, these activities could be conveniently filtered out using basic detection methods, such as those based on accelerometer data.

Spotting Temporal Resolution

BS spotting requires a temporal resolution in the millisecond scale to detect very short events (<100 ms). Previous studies have attempted to maximize the temporal resolution by minimizing the sliding window applied to inspect the audio data (eg, [9,41]). When recording BSs with wearable devices, however, the temporal sparsity of BS events could increase as a result of sensor displacements or noise artifacts. As the acoustic context decreases with sliding window duration, false-positive cases may increase, thus limiting spotting performance. Our multiscale approach can analyze continuous recordings with a temporal resolution of 25 ms while retaining 10 seconds of acoustic context in the spotting by the audio segment S_i .

Comparison Between Healthy and Patients With IBD

Our DNN model EffUNet can detect BSs with a median precision of 80% and a median recall of 73%. Although the median F_1 -score for healthy participants and patients was the same, BS spotting was more challenging in patients, as the difference in the median recall of 66% versus 75% indicates (Table 2). The patient group included individuals with different IBDs and varying levels of inflammation activity, which may explain the greater variability of acoustic patterns in BSs, compared with the healthy group. Our results warrant further data recordings from patients with IBD. A nested validation set could be used to analyze model hyperparameters. In this work, however, our focus was to maximize trainset size and minimize model bias. Thus, we used LOPO CV without early stopping criteria during training and other training parameters were chosen according to the Pretraining, Sampling, Labeling, and Aggregation pipeline [24]. The F_1 -score was above 60% for all patients, except for 1 outlier (P22; see Figures 3 and 4), where the performance of EffUNet dropped to approximately 50% as a result of the reduced recall. As the F_1 -score of EffUNet showed an IQR of 0.14 across all patients, we attribute the performance drop for P22 to a reduced recording quality: Less than 100 annotated BS events across all channels were documented (BS ratio=0.0028). Analysis of the false-positive rate showed that EffUNet spotted events that were marked by the raters as tentative BSs because of their noisy patterns. If tentative BSs had been included in the evaluation for P22, the model's precision would have increased from 52% to 83%. However, tentative events were not considered in our analysis because of their noisy acoustic patterns and were labeled as NBSs. Thus, assigning tentative annotations to NBSs, that is, the NULL class, may have increased overall insertion errors and thus contributed to a conservative performance estimation. Further investigations on data collection and preprocessing, for example, adaptive noise filtering [12], could improve signal quality and consequently spotting performance as well.

Spotting Performance Under Different Noise Conditions

Compared with other studies, where BSs were recorded using a skin-taped sensor [42], our work used garment-embedded microphones. Continuous data collection with wearable devices could further decrease the signal amplitude as a result of accidental sensor displacement and motion artifacts. Nevertheless, our approach can spot BSs with recall greater than 73% regardless of the noise level (Figure 3). In addition, our precision over SNR analysis demonstrated that our model was robust against background noise, as more than 70% of predictions corresponded to ground truth events even when SNR=-20 dB. In particular, for low SNR conditions, empirical threshold-based BS detection methods could fail, as reported, for example, by Sato et al [41]. We attribute EffUNet's reduced number of false positives to the encoder's pretraining on AudioSet. EfficientNet-B2 was originally trained to detect sound events in audio clips of duration $\delta=10$ seconds [24]. In the experiments on AudioSet, EfficientNet-B2 achieved an average precision of ≈ 0.44 for classifying 527 sound classes. The pretraining on a large variety of noise sources could have



improved the modeling of the NBS class, and thus, model precision. However, AudioSet does not provide strong audio labels (ie, event onset/offset), and therefore, no pretraining could be applied to the EffUNet decoder.

Spotting Timing Errors

Previous studies on sound event detection (eg, [36]) highlighted that common pattern recognition evaluation metrics are insufficient to describe error types in continuous data. For instance, a model could return a fragmented prediction of a ground truth event or could recognize multiple events in 1 prediction. Previous work on BS spotting rarely analyzed detection errors besides the false-positive rate. As missed BSs will decrease the number of BS events per unit time, diagnostic approaches based on BS event count thresholding (eg, [8]) may fail to identify IBD. Furthermore, timing errors may affect the diagnosis of gastrointestinal disorders. Fragmentation or merge errors could alter natural BS acoustic characteristics (ie, spectral and temporal features), which were explored in previous studies to classify digestive dysfunctions (eg, IBD [43]) or to detect digestive events (eg, migrating motor complex [44]). Our analysis of detection errors (Table 3) showed that the performance of the EffUNet model was mainly impacted by insertions (ie, false positives) and deletions (ie, missed BSs). In 136 hours of audio data, 19.56 minutes of background noise were wrongly detected as BSs, because of either insertion or overfill errors. Insertion errors were largest in the healthy group (13.28 minutes out of 84 hours of audio data), probably because of the larger group size compared with the patient group, and consequently more variable background noise. Of the 1.22 hours of audio annotated as BSs, 18.56 minutes were not recognized because of deletion and underfill errors. Deletion errors were largest in the patient group (5.15 minutes out of the annotated 20.71 minutes), as confirmed by the lower recall compared with the healthy group (66% vs 75%). Nevertheless, our training loss (Equation 5) could minimize fragmentation and merge errors (ie, EffUNet returned prediction onset/offset according to our annotation approach). Overall, overfill and underfill errors were 4.01 and 5.56 minutes, respectively, and peaked in the healthy group. As BS annotations were converted to a binary mask to train EffUNet (Equation 4), we hypothesize that further improvements on the input data preprocessing (eg, spectrogram sliding window size γ and stride length σ) could improve the detection temporal resolution, thus minimizing overfill and underfill errors.

Spotting Performance Over BS Event Duration

As described by previous studies [17,18], BSs present acoustic patterns of variable duration. In our study, BS length varied from 18 ms to 6.29 seconds, which created a challenging spotting task. To detect very short events (ie, <100 ms), previous work typically used a sliding window with a duration no more than the BS length (eg, [15,41]). Because of the constrained data included in the window, a spotter thus has limited information available to spot BSs. As for human experts, spotting performance could decrease when context information from the surrounding audio scene diminishes. By contrast, a larger sliding window could decrease temporal resolution, and thus yield coarse event onset/offset prediction (eg, [45]). In our

work, we propose a multiscale approach: The data stream is first split into 10-second audio segments S_i , and for each S_i , a binary mask M_i is generated. While analyzing S_i with duration = 10 seconds at a time, EffUNet can detect BSs with a temporal resolution of 25 ms. Our approach could be potentially applied to other spotting tasks where events have a varying duration (eg, gesture recognition [46]). Our analysis on recall versus event duration e_j showed a decrease in recall for long BSs (>1 second; Figure 3). As, in our analysis, the duration threshold θ_D was applied to events e_j as well as predictions , we hypothesize that predictions for longer events could have been affected by underfill or fragmentation errors and, consequently, filtered out during the analysis. Long BS events have been previously described as a sequence of single and multiple bursts interrupted by silence periods. In our BS annotation approach (see the “BS Annotation” section), silence periods between consecutive bursts of a maximum 100 ms were accepted. Therefore, some parts of a long BS (>1 second) could have been rejected by EffUNet as noise. Additional postprocessing on spotting results (eg, merging nearby detected events ) or alternative loss functions (eg, based on dice loss [29]) could improve the retrieval of long BSs (>1 second). For instance, scaling factors could be explored when combining the cross-entropy loss with the dice loss during EffUNet training. In this work, no weighting was applied when calculating the loss during backpropagation (Equation 5).

Spotting Performance for Different Event Rates

Previous studies have already introduced shallower DNNs than EffUNet to spot BSs in the data streams, demonstrating promising results (eg, [9,15]). However, the previously reported models were trained and tested on limited, partially selected data subsets, often with a BS ratio of 0.50 (ie, class balance between BSs and NBSs; Table 4). When spotting BS events in continuous recordings that are collected in daily settings, however, a substantial BS versus NULL class imbalance must be expected (see Figures 2-4). Algorithm evaluations on a balanced data set could therefore overestimate performance for a BS ratio $\ll 0.50$. The difference can be seen between Wang et al’s [9] original report and the analysis of their CNN on our data set. However, Wang et al’s [9] CNN was designed for balanced BS detection, which limits a direct comparison with EffUNet in our study. In our data set, BSs were highly sparse, with BS ratios less than 0.035 across all participants, corresponding to event rates of approximately 100-300 events/hour. With event sparsity, the spotting challenge increases [16], especially when training and evaluating the spotter on different class imbalances. Despite the high class imbalance, however, EffUNet could retrieve BSs with a recall of 73% and a precision of 72%. By contrast, Ficek et al [15] reported a precision of 83% at a BS ratio of 0.15, but yielded a precision of 58% for a BS ratio of 0.0246. If our data set had a BS ratio of 0.15 (approximately 1400 events/hour), the precision of our DNN would reach an estimated 92% (Figure 3).

Spotting Performance Over Sensor Location

We compared PR metrics across different sensor locations (Figure 4). EffUNet yielded comparable median recall across

all locations, although recall IQR was largest at the large intestine. As performance median and IQR were comparable for all sensors, we hypothesized that the higher recall IQR could be due to abnormal BS patterns that are more likely to occur in the large intestine. Although the sensor data from the large intestine were annotated only for a subset of the study participants, performance was not affected by the data imbalance across channels.

Limitations

While our wearable prototype design allowed us to capture BSs with multiple sensors, no sensor fusion was applied as in [12,37]. Previous experiments by Ranta et al [47] showed that the abdomen can be acoustically modeled as an absorbent material, and thus BS intensity depended on sensor distance. Our data annotation confirmed that not all BS events were captured by all channels. Given past analyses on abdominal sound propagation, we decided to minimize the model complexity and designed a single-channel spotting model. Further investigations on abdominal sound propagation may improve BS source localization and estimate the relationship between BSs and bowel movements. For instance, the EffUNet architecture could be extended to analyze multichannel recordings and locate BS sources. Preliminary studies on BS source location [47], however, suggested that further basic analyses on sound propagation in the abdominal cavity are needed. BS source localization would be beneficial for patients with IBD, for example, to locate inflammation sites noninvasively based on abnormal BS patterns and related digital biomarkers.

Although our analysis compared EffUNet spotting performance across study populations with different gastrointestinal conditions (healthy volunteers and patients with IBD), the impact of false positives on BS-based clinical gastrointestinal assessment was not evaluated. Future studies should investigate methods for digestive disorder recognition. Based on the spotting method proposed in this work, a fully automated and noninvasive approach for digestive disorder analysis may be feasible.

The usability and comfort of the wearable prototype were not analyzed in depth in this work. A full user comfort study is beyond the scope of this analysis. Nevertheless, we carefully considered user comfort during the T-shirt implementation, especially because our investigation involved patients and the longest recordings analyzed thus far. While participants did not express complaints about the wearing comfort, which confirms the efficacy of our approach, further research could explore design optimizations supported by a focused wearability assessment.

Our study encompassed 2 hours of annotated audio from 2-3 channels for each of the 27 participants, resulting in 136 hours of labeled data. To the best of our knowledge, our data set is the largest annotated BS data set reported to date. Our results justify further studies with even larger data sets. Although BSs were recorded across different digestive phases, further investigations should include data collected from extended monitoring periods, such as over multiple days, and in less constrained settings, such as in a home setting. This approach would better facilitate the correlation of physiological digestive

processes with BS acoustic properties and enable the investigation of noise effects.

Comparison With Prior Work

The goal of our work was to design and evaluate an approach to deal with imbalanced BS data gathered from body-worn microphone sensors. Therefore, when comparing EffUNet with previous studies, we focused our analysis on models with similar scope (ie, BS spotting with maximized temporal spotting resolution). As reported by Ficek et al [15], data scarcity and missing annotations as a result of labor-intensive audio recording inspection are well-known challenges in the field of BS analysis. Open BS data sets are inexistent so far, which may be due to privacy concerns associated with raw audio recordings. Moreover, BSs are often collected in different recording settings, using various wearable devices and following varying recording protocols. Consequently, benchmarking our EffUNet directly against past BS spotting models is challenging. We excluded architectures that are similar to EffUNet but were not designed for BS detection (eg, UNet [19]). According to Table 4, only 3 methods achieved spotting resolution below 100 ms. The CRNN by Ficek et al [15] could not be scaled to our much larger data set. Further, the CNN proposed by Kutsumi et al [38] could not be included in our comparison, because it lacked methodology details that are required to reimplement the model (eg, training optimizer and dropout layer parameters). Nevertheless, we reimplemented the CNN by Wang et al [9]. Compared with EffUNet, the CNN is a shallower network, with approximately 62,000 parameters versus approximately 18.1 million EffUNet. Our model outperformed the CNN by Wang et al [9] not only in temporal spotting resolution (25 ms of EffUNet vs 60 ms of

the CNN), but also in spotting performance on highly imbalanced data (80% median precision of EffUNet vs 5% median precision of the CNN). Using EfficientNet for our model encoder allowed us to leverage pretraining to increase model robustness against noise, as shown in [14]. Further work may investigate whether less complex models than EffUNet could be optimized for natural, imbalanced data. Moreover, our analysis of related work (Table 4) provides a comparison of recent advances in BS spotting, showing how our work outperforms previous studies in dealing with data imbalance and temporal spotting resolution.

Conclusions

We presented a multiscale BS spotting model based on the EffUNet architecture, to detect BSs in continuous audio data streams. AudioSet pretraining was applied to the EffUNet encoder to improve model robustness against noise. We evaluated our model using 136 hours of audio data collected from 18 healthy participants and 9 patients with IBD. Our experiments demonstrated that EffUNet can detect BSs with a median F_1 -score of 73% in recordings where BS events were highly sparse (BS ratio of 0.0089). With EffUNet, BSs of varying durations and under different noise conditions could be identified with a precision of 72%. Our EffUNet analysis surpassed previous approaches not only in terms of evaluation data size and temporal sparsity of BS events but also achieved one of the highest temporal resolutions. Using our approach, future analyses of BSs obtained from wearable abdominal monitoring systems could be automated without requiring manual audio data annotation.

Acknowledgments

The authors gratefully acknowledge the HPC resources provided by the Erlangen National High Performance Computing Centre (NHR@FAU) of the Friedrich-Alexander-Universität Erlangen-Nürnberg under NHR project b131dc. NHR funding was provided by Federal and Bavarian state authorities. NHR@FAU hardware was partially funded by the German Research Foundation (DFG; grant 440719683).

Conflicts of Interest

None declared.

References

1. Grønlund D, Poulsen JL, Sandberg TH, Olesen AE, Madzak A, Krogh K, et al. Established and emerging methods for assessment of small and large intestinal motility. *Neurogastroenterol Motil* 2017 Jul 13;29(7):e13008. [doi: [10.1111/nmo.13008](https://doi.org/10.1111/nmo.13008)] [Medline: [28086261](https://pubmed.ncbi.nlm.nih.gov/28086261/)]
2. Dhyani M, Joshi N, Bemelman W, Gee M, Yajnik V, D'Hoore A, et al. Challenges in IBD research: novel technologies. *Inflamm Bowel Dis* 2019 May 16;25(Suppl 2):S24-S30 [FREE Full text] [doi: [10.1093/ibd/izz077](https://doi.org/10.1093/ibd/izz077)] [Medline: [31095703](https://pubmed.ncbi.nlm.nih.gov/31095703/)]
3. Fritz D, Weilitz P. Abdominal assessment. *Home Healthcare Now* 2016;34(3):151-155. [doi: [10.1097/nhh.0000000000000364](https://doi.org/10.1097/nhh.0000000000000364)]
4. Stevens N. Auscultation of the abdomen. *N Engl J Med* 1936 Jul 02;215(1):22-26. [doi: [10.1056/NEJM193607022150102](https://doi.org/10.1056/NEJM193607022150102)]
5. Li B, Wang JR, Ma YL. Bowel sounds and monitoring gastrointestinal motility in critically ill patients. *Clin Nurse Spec* 2012;26(1):29-34. [doi: [10.1097/NUR.0b013e31823bfab8](https://doi.org/10.1097/NUR.0b013e31823bfab8)] [Medline: [22146271](https://pubmed.ncbi.nlm.nih.gov/22146271/)]
6. Reuben A. Examination of the abdomen. *Clin Liver Dis (Hoboken)* 2016 Jun;7(6):143-150 [FREE Full text] [doi: [10.1002/cld.556](https://doi.org/10.1002/cld.556)] [Medline: [31041050](https://pubmed.ncbi.nlm.nih.gov/31041050/)]
7. Ranta R, Louis-Dorr V, Heinrich CH, Wolf D, Guillemin F. Principal component analysis and interpretation of bowel sounds. New York, NY: IEEE; 2004 Presented at: The 26th Annual International Conference of the IEEE Engineering in

- Medicine and Biology Society; September 1-5, 2004; San Francisco, CA, USA p. 227-230. [doi: [10.1109/iembs.2004.1403133](https://doi.org/10.1109/iembs.2004.1403133)]
8. Craine B, Silpa M, O'Toole C. Enterotachogram analysis to distinguish irritable bowel syndrome from Crohn's disease. *Dig Dis Sci* 2001 Sep;46(9):1974-1979. [doi: [10.1023/a:1010651602095](https://doi.org/10.1023/a:1010651602095)] [Medline: [11575452](https://pubmed.ncbi.nlm.nih.gov/11575452/)]
 9. Wang N, Testa A, Marshall BJ. Development of a bowel sound detector adapted to demonstrate the effect of food intake. *Biomed Eng Online* 2022 Jan 04;21(1):1 [FREE Full text] [doi: [10.1186/s12938-021-00969-2](https://doi.org/10.1186/s12938-021-00969-2)] [Medline: [34983542](https://pubmed.ncbi.nlm.nih.gov/34983542/)]
 10. Yao C, Tai W. P345 Application of bowel sound computational analysis in inflammatory bowel disease. *J Crohns Colitis* 2024;18(Supplement_1):i741. [doi: [10.1093/ecco-jcc/jjad212.0475](https://doi.org/10.1093/ecco-jcc/jjad212.0475)]
 11. Baid H. A critical review of auscultating bowel sounds. *Br J Nurs* 2009;18(18):1125-1129. [doi: [10.12968/bjon.2009.18.18.44555](https://doi.org/10.12968/bjon.2009.18.18.44555)] [Medline: [19966732](https://pubmed.ncbi.nlm.nih.gov/19966732/)]
 12. Wang G, Yang Y, Chen S, Fu J, Wu D, Yang A, et al. Flexible dual-channel digital auscultation patch with active noise reduction for bowel sound monitoring and application. *IEEE J Biomed Health Inform* 2022 Jul;26(7):2951-2962. [doi: [10.1109/JBHI.2022.3151927](https://doi.org/10.1109/JBHI.2022.3151927)] [Medline: [35171784](https://pubmed.ncbi.nlm.nih.gov/35171784/)]
 13. Liu J, Yin Y, Jiang H, Kan H, Zhang Z, Chen P, et al. Bowel sound detection based on MFCC feature and LSTM neural network. New York, NY: IEEE; 2018 Presented at: 2018 IEEE Biomedical Circuits and Systems Conference (BioCAS); October 17-19, 2018; Cleveland, OH p. 1-4. [doi: [10.1109/biocas.2018.8584723](https://doi.org/10.1109/biocas.2018.8584723)]
 14. Baronetto A, Graf LS, Fischer S, Neurath MF, Amft O. Segment-based spotting of bowel sounds using pretrained models in continuous data streams. *IEEE J Biomed Health Inform* 2023 Jul;27(7):3164-3174. [doi: [10.1109/JBHI.2023.3269910](https://doi.org/10.1109/JBHI.2023.3269910)] [Medline: [37155392](https://pubmed.ncbi.nlm.nih.gov/37155392/)]
 15. Ficek J, Radzikowski K, Nowak JK, Yoshie O, Walkowiak J, Nowak R. Analysis of gastrointestinal acoustic activity using deep neural networks. *Sensors (Basel)* 2021 Nov 16;21(22):7602 [FREE Full text] [doi: [10.3390/s21227602](https://doi.org/10.3390/s21227602)] [Medline: [34833679](https://pubmed.ncbi.nlm.nih.gov/34833679/)]
 16. Amft O. Adaptive activity spotting based on event rates. New York, NY: IEEE; 2010 Presented at: 2010 IEEE International Conference on Sensor Networks, Ubiquitous, and Trustworthy Computing; June 7-9, 2010; Newport Beach, CA p. 169. [doi: [10.1109/sutc.2010.63](https://doi.org/10.1109/sutc.2010.63)]
 17. Dimoulas CA, Papanikolaou GV, Petridis V. Pattern classification and audiovisual content management techniques using hybrid expert systems: a video-assisted bioacoustics application in abdominal sounds pattern analysis. *Expert Systems with Applications* 2011 Sep;38(10):13082-13093. [doi: [10.1016/j.eswa.2011.04.115](https://doi.org/10.1016/j.eswa.2011.04.115)]
 18. Du X, Allwood G, Webberley KM, Osseiran A, Wan W, Volikova A, et al. A mathematical model of bowel sound generation. *J Acoust Soc Am* 2018 Dec;144(6):EL485. [doi: [10.1121/1.5080528](https://doi.org/10.1121/1.5080528)] [Medline: [30599659](https://pubmed.ncbi.nlm.nih.gov/30599659/)]
 19. Ronneberger O, Fischer P, Brox T. U-Net: convolutional networks for biomedical image segmentation. Cham, Switzerland: Springer; 2015 Presented at: The International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI 2015); October 5-9, 2015; Munich, Germany p. 234-241. [doi: [10.1007/978-3-319-24574-4_28](https://doi.org/10.1007/978-3-319-24574-4_28)]
 20. Tan M, Le Q. EfficientNet: rethinking model scaling for convolutional neural networks. In: *Proceedings of the 36th International Conference on Machine Learning*. 2019 Presented at: 36th International Conference on Machine Learning; June 9-15, 2019; Long Beach, CA. [doi: [10.1007/978-1-4842-6168-2_10](https://doi.org/10.1007/978-1-4842-6168-2_10)]
 21. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016 Presented at: IEEE Conference on Computer Vision and Pattern Recognition; June 27-30, 2016; Las Vegas, NV p. 03385. [doi: [10.1109/cvpr.2016.90](https://doi.org/10.1109/cvpr.2016.90)]
 22. Stoller D, Ewert S, Dixon S. Wave-U-Net: A Multi-Scale Neural Network for End-to-End Audio Source Separation. *arXiv Preprint posted online June 8, 2018*. [doi: [10.48550/arXiv.1806.03185](https://doi.org/10.48550/arXiv.1806.03185)]
 23. Baheti B, Innani S, Gajre S, Talbar S. Eff-UNet: a novel architecture for semantic segmentation in unstructured environment. 2020 Presented at: IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops; June 14-19, 2020; Seattle, WA p. 358-359. [doi: [10.1109/cvprw50498.2020.00187](https://doi.org/10.1109/cvprw50498.2020.00187)]
 24. Gong Y, Chung YA, Glass J. PSLA: improving audio tagging with pretraining, sampling, labeling, and aggregation. *IEEE/ACM Trans Audio Speech Lang Process* 2021;29:3292-3306. [doi: [10.1109/TASLP.2021.3120633](https://doi.org/10.1109/TASLP.2021.3120633)]
 25. Gemmeke J, Ellis D, Freedman D, Jansen A, Lawrence W, Moore R, et al. Audio Set: an ontology and human-labeled dataset for audio events. New York, NY: IEEE; 2017 Presented at: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); March 5-9, 2017; New Orleans, LA p. 776-780. [doi: [10.1109/icassp.2017.7952261](https://doi.org/10.1109/icassp.2017.7952261)]
 26. He K, Zhang X, Ren S, Sun J. Delving deep into rectifiers: surpassing human-level performance on ImageNet classification. New York, NY: IEEE; 2015 Presented at: The IEEE International Conference on Computer Vision (ICCV); December 7-13, 2015; Santiago, Chile. [doi: [10.1109/iccv.2015.123](https://doi.org/10.1109/iccv.2015.123)]
 27. Kingma DP, Ba J. Adam: a method for stochastic optimization. *arXiv Preprint posted online on January 30, 2017* [FREE Full text] [doi: [10.48550/arXiv.1412.6980](https://doi.org/10.48550/arXiv.1412.6980)]
 28. Bishop C. *Pattern Recognition and Machine Learning*. New York, NY: Springer; 2006:5-43.
 29. Sørensen T. *A Method of Establishing Groups of Equal Amplitude in Plant Sociology Based on Similarity of Species Content and Its Application to Analyses of the Vegetation on Danish Commons*. Copenhagen, Denmark: Munksgaard in Komm; 1948.

30. Park DS, Chan W, Zhang Y, Chiu CC, Zoph B, Cubuk ED, et al. Specaugment: a simple data augmentation method for automatic speech recognition. arXiv Preprint posted online on December 3, 2019 [FREE Full text] [doi: [10.21437/Interspeech.2019-2680](https://doi.org/10.21437/Interspeech.2019-2680)]
31. Baronetto A, Graf L, Fischer S, Neurath M, Amft O. GastroDigitalShirt: a smart shirt for digestion acoustics monitoring. New York, NY: Association for Computing Machinery; 2020 Presented at: The 2020 ACM International Symposium on Wearable Computers; September 12-16, 2020; Virtual p. 17-21. [doi: [10.1145/3410531.3414297](https://doi.org/10.1145/3410531.3414297)]
32. de Oliveira EP, Burini RC. The impact of physical exercise on the gastrointestinal tract. *Curr Opin Clin Nutr Metab Care* 2009 Sep;12(5):533-538. [doi: [10.1097/MCO.0b013e32832e6776](https://doi.org/10.1097/MCO.0b013e32832e6776)] [Medline: [19535976](https://pubmed.ncbi.nlm.nih.gov/19535976/)]
33. Knight J, Baber C, Schwirtz A, Bristow H. The comfort assessment of wearable computers. 2002 Presented at: 2002 International Symposium on Wearable Computers; October 7-10, 2002; Seattle, WA. [doi: [10.1109/iswc.2002.1167220](https://doi.org/10.1109/iswc.2002.1167220)]
34. Fleiss J, Levin B, Paik M. *Statistical Methods for Rates and Proportions*. Hoboken, NJ: John Wiley & Sons; 2013.
35. Mesaros A, Heittola T, Virtanen T. Metrics for polyphonic sound event detection. *Applied Sciences* 2016 May 25;6(6):162. [doi: [10.3390/app6060162](https://doi.org/10.3390/app6060162)]
36. Ward JA, Lukowicz P, Gellersen HW. Performance metrics for activity recognition. *ACM Trans Intell Syst Technol* 2011 Jan 24;2(1):1-23. [doi: [10.1145/1889681.1889687](https://doi.org/10.1145/1889681.1889687)]
37. Bilonis I, Apostolidis G, Charisis V, Liatsos C, Hadjileontiadis L. Non-invasive detection of bowel sounds in real-life settings using spectrogram zeros and autoencoding. New York, NY: IEEE; 2021 Presented at: 2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC); November 1-5, 2021; Mexico City, Mexico p. 915-919. [doi: [10.1109/embc46164.2021.9630783](https://doi.org/10.1109/embc46164.2021.9630783)]
38. Kutsumi Y, Kanegawa N, Zeida M, Matsubara H, Murayama N. Automated bowel sound and motility analysis with CNN using a smartphone. *Sensors (Basel)* 2022 Dec 30;23(1):407 [FREE Full text] [doi: [10.3390/s23010407](https://doi.org/10.3390/s23010407)] [Medline: [36617005](https://pubmed.ncbi.nlm.nih.gov/36617005/)]
39. Zhao Z, Li F, Xie Y, Wu Y, Wang Y. BSMonitor: noise-resistant bowel sound monitoring via earphones. *IEEE Trans on Mobile Comput* 2024 Apr;23(4):3213-3227. [doi: [10.1109/TMC.2023.3270926](https://doi.org/10.1109/TMC.2023.3270926)]
40. Burne L, Sitaula C, Priyadarshi A, Tracy M, Kavehei O, Hinder M, et al. Ensemble approach on deep and handcrafted features for neonatal bowel sound detection. *IEEE J Biomed Health Inform* 2023 Jun;27(6):2603-2613. [doi: [10.1109/JBHI.2022.3217559](https://doi.org/10.1109/JBHI.2022.3217559)] [Medline: [36301790](https://pubmed.ncbi.nlm.nih.gov/36301790/)]
41. Sato R, Emoto T, Gojima Y, Akutagawa M. Automatic bowel motility evaluation technique for noncontact sound recordings. *Applied Sciences* 2018 Jun 19;8(6):999. [doi: [10.3390/app8060999](https://doi.org/10.3390/app8060999)]
42. Dimoulas C, Kalliris G, Papanikolaou G, Kalampakas A. Long-term signal detection, segmentation and summarization using wavelets and fractal dimension: a bioacoustics application in gastrointestinal-motility monitoring. *Comput Biol Med* 2007 Apr;37(4):438-462. [doi: [10.1016/j.compbiomed.2006.08.013](https://doi.org/10.1016/j.compbiomed.2006.08.013)] [Medline: [17026978](https://pubmed.ncbi.nlm.nih.gov/17026978/)]
43. Du X, Allwood G, Webberley KM, Inderjeeth AJ, Osseiran A, Marshall BJ. Noninvasive diagnosis of irritable bowel syndrome via bowel sound features: proof of concept. *Clin Transl Gastroenterol* 2019 Mar;10(3):e00017 [FREE Full text] [doi: [10.14309/ctg.000000000000017](https://doi.org/10.14309/ctg.000000000000017)] [Medline: [30908308](https://pubmed.ncbi.nlm.nih.gov/30908308/)]
44. Kölle K, Fougner AL, Ellingsen R, Carlsen SM, Stavdahl O. Feasibility of early meal detection based on abdominal sound. *IEEE J Transl Eng Health Med* 2019;7:3300212 [FREE Full text] [doi: [10.1109/JTEHM.2019.2940218](https://doi.org/10.1109/JTEHM.2019.2940218)] [Medline: [32309058](https://pubmed.ncbi.nlm.nih.gov/32309058/)]
45. Zhao K, Feng S, Jiang H, Wang Z, Chen P, Zhu B, et al. A Binarized CNN-based Bowel Sound Recognition Algorithm with Time-domain Histogram Features for Wearable Healthcare Systems. *IEEE Trans Circuits Syst II Express Briefs* 2022;71:1. [doi: [10.1109/TCSII.2021.3097069](https://doi.org/10.1109/TCSII.2021.3097069)]
46. Benitez-Garcia G, Haris M, Tsuda Y, Ukita N. Finger gesture spotting from long sequences based on multi-stream recurrent neural networks. *Sensors (Basel)* 2020 Jan 18;20(2):528 [FREE Full text] [doi: [10.3390/s20020528](https://doi.org/10.3390/s20020528)] [Medline: [31963623](https://pubmed.ncbi.nlm.nih.gov/31963623/)]
47. Ranta R, Louis-Dorr V, Heinrich C, Wolf D, Guillemin F. Towards an acoustic map of abdominal activity. New York, NY: IEEE; 2003 Presented at: The 25th Annual International Conference of the IEEE Engineering in Medicine and Biology Society; September 17-21, 2003; Cancun, Mexico p. 2769-2772. [doi: [10.1109/iembs.2003.1280491](https://doi.org/10.1109/iembs.2003.1280491)]

Abbreviations

- BS:** bowel sound
- CNN:** convolutional neural network
- CRNN:** convolutional recurrent neural network
- CV:** cross-validation
- DNN:** deep neural network
- EffUNet:** Efficient-U-Net
- IBD:** inflammatory bowel disease
- LOPO:** leave-one-participant-out
- NBS:** nonbowel sound
- PR:** precision and recall metrics
- SNR:** signal-to-noise ratio
- up-conv:** transposed convolution

Edited by K El Emam, B Malin; submitted 25.07.23; peer-reviewed by K Zhang, G Lim, A Staffini; comments to author 17.02.24; revised version received 29.03.24; accepted 24.04.24; published 10.07.24.

Please cite as:

Baronetto A, Graf L, Fischer S, Neurath MF, Amft O

Multiscale Bowel Sound Event Spotting in Highly Imbalanced Wearable Monitoring Data: Algorithm Development and Validation Study

JMIR AI 2024;3:e51118

URL: <https://ai.jmir.org/2024/1/e51118>

doi: [10.2196/51118](https://doi.org/10.2196/51118)

PMID: [38985504](https://pubmed.ncbi.nlm.nih.gov/38985504/)

©Annalisa Baronetto, Luisa Graf, Sarah Fischer, Markus F Neurath, Oliver Amft. Originally published in JMIR AI (<https://ai.jmir.org>), 10.07.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR AI, is properly cited. The complete bibliographic information, a link to the original publication on <https://www.ai.jmir.org/>, as well as this copyright and license information must be included.

Original Paper

Leveraging Machine Learning to Develop Digital Engagement Phenotypes of Users in a Digital Diabetes Prevention Program: Evaluation Study

Danissa V Rodriguez¹, PhD; Ji Chen¹, PhD; Ratnalekha V N Viswanadham¹, PhD; Katharine Lawrence^{1,2}, MPH, MD; Devin Mann^{1,2}, MS, MD

¹New York University Grossman School of Medicine, New York, NY, United States

²New York University Langone Health, New York, NY, United States

Corresponding Author:

Danissa V Rodriguez, PhD

New York University Grossman School of Medicine

227 E 30th St, 6th Floor

New York, NY, 10016

United States

Phone: 1 914 320 7655

Email: danissa.rodriguez@nyulangone.org

Abstract

Background: Digital diabetes prevention programs (dDPPs) are effective “digital prescriptions” but have high attrition rates and program noncompletion. To address this, we developed a personalized automatic messaging system (PAMS) that leverages SMS text messaging and data integration into clinical workflows to increase dDPP engagement via enhanced patient-provider communication. Preliminary data showed positive results. However, further investigation is needed to determine how to optimize the tailoring of support technology such as PAMS based on a user’s preferences to boost their dDPP engagement.

Objective: This study evaluates leveraging machine learning (ML) to develop digital engagement phenotypes of dDPP users and assess ML’s accuracy in predicting engagement with dDPP activities. This research will be used in a PAMS optimization process to improve PAMS personalization by incorporating engagement prediction and digital phenotyping. This study aims (1) to prove the feasibility of using dDPP user-collected data to build an ML model that predicts engagement and contributes to identifying digital engagement phenotypes, (2) to describe methods for developing ML models with dDPP data sets and present preliminary results, and (3) to present preliminary data on user profiling based on ML model outputs.

Methods: Using the gradient-boosted forest model, we predicted engagement in 4 dDPP individual activities (physical activity, lessons, social activity, and weigh-ins) and general activity (engagement in any activity) based on previous short- and long-term activity in the app. The area under the receiver operating characteristic curve, the area under the precision-recall curve, and the Brier score metrics determined the performance of the model. Shapley values reflected the feature importance of the models and determined what variables informed user profiling through latent profile analysis.

Results: We developed 2 models using weekly and daily DPP data sets (328,821 and 704,242 records, respectively), which yielded predictive accuracies above 90%. Although both models were highly accurate, the daily model better fitted our research plan because it predicted daily changes in individual activities, which was crucial for creating the “digital phenotypes.” To better understand the variables contributing to the model predictor, we calculated the Shapley values for both models to identify the features with the highest contribution to model fit; engagement with any activity in the dDPP in the last 7 days had the most predictive power. We profiled users with latent profile analysis after 2 weeks of engagement (Bayesian information criterion=-3222.46) with the dDPP and identified 6 profiles of users, including those with high engagement, minimal engagement, and attrition.

Conclusions: Preliminary results demonstrate that applying ML methods with predicting power is an acceptable mechanism to tailor and optimize messaging interventions to support patient engagement and adherence to digital prescriptions. The results enable future optimization of our existing messaging platform and expansion of this methodology to other clinical domains.

Trial Registration: ClinicalTrials.gov NCT04773834; <https://www.clinicaltrials.gov/ct2/show/NCT04773834>

International Registered Report Identifier (IRRID): RR2-10.2196/26750

KEYWORDS

machine learning; digital health; diabetes; mobile health; messaging platforms; user engagement; patient behavior; digital diabetes prevention programs; digital phenotypes; digital prescription; users; prevention; evaluation study; communication; support; engagement; phenotypes; digital health intervention; chronic disease management

Introduction

Over 80 million US adults have prediabetes, a metabolic condition that places individuals at risk for progression to type 2 diabetes and its related complications [1]. Evidence-based strategies for diabetes prevention have primarily focused on nonpharmacologic interventions such as diabetes prevention programs (DPPs), which are comprehensive behavior change curricula concentrating on physical activity and dietary modification. Such programs can be as effective as medication in preventing the progression of diabetes in at-risk populations [2]. Increasingly, DPP behavioral curricula have been adapted to digital platforms (digital DPPs [dDPPs]), which have demonstrated comparable effectiveness in achieving weight loss, hemoglobin A_{1c} reduction, and other critical diabetes-related health outcomes while offering improvements in accessibility, convenience, and personalization [3]. Yet, limited patient engagement with digital interventions presents a significant barrier to translating evidence-based digital behavioral interventions such as the dDPP into pragmatic, scalable solutions [4-8].

To address this critical patient engagement issue, various technologies and interventions have been developed to provide targeted support to patients using digital health apps to improve engagement and sustained use [9]. Potential solutions include mobile-based feedback and reminder tools, app-based coaching, social networking, and gamification. More recent strategies have also leveraged machine learning (ML) and big data analytics to deploy more advanced tools, such as engagement algorithms and artificial intelligence (AI)-driven chatbots. ML solutions can provide (1) more nuanced patient segmentation or phenotyping; (2) more precise, tailored interventions, with enhanced ability to respond dynamically to changes in individual trends; and (3) improved resource alignment by intervention implementers, as automated processes (eg, chatbots) can free up human capital for more appropriate tasks [10]. Moreover, AI-driven chatbots (AI chatbots), conversational agents that mimic human interaction through written, oral, and visual communication channels with a user [1,2], have demonstrated efficacy in health-behavior change interventions among a large and diverse population [3-6,11-13].

Prior work from this team involved developing a personalized automatic messaging system (PAMS) that leveraged an evidence-based engagement algorithm to deliver tailored behavior change theory-supported SMS text messaging to support users engaging with a commercial app-based dDPP.

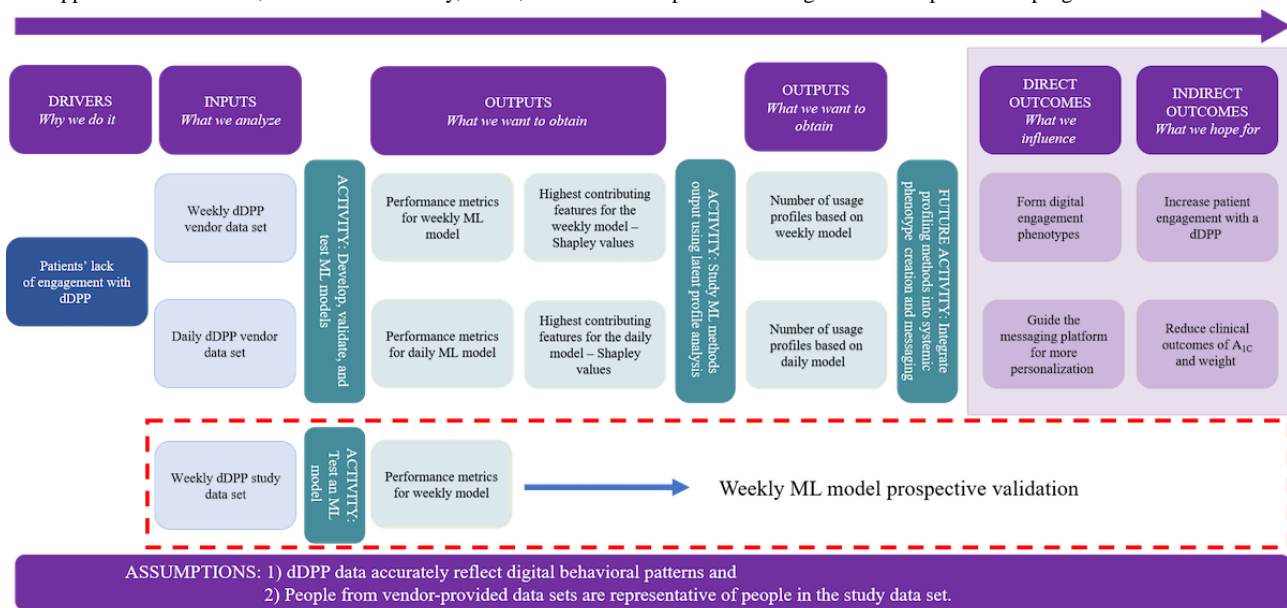
The study returned promising results compared with average users, demonstrating engagement in various dDPP features (eg, weight tracking and physical activity logins) [12]. To expand on the previous investigation, improved features of the next generation of PAMS include an ML-based patient engagement prediction algorithm to identify dDPP digital engagement phenotypes and to guide and further personalize the messaging intervention. This paper describes the ML model designed to predict characteristics and behavioral patterns of dDPP user types (eg, those highly engaged with exercise but not uploading the meals or those messaging their coach but not participating in weigh-ins) based on their activity patterns within a dDPP app, with a particular focus on motivating users at risk for low engagement and nonengagement with the dDPP (ie, patient digital engagement phenotypes).

Methods

Overview

The logic diagram in Figure 1 illustrates, from left to right, the overall framework for optimizing patient engagement with a dDPP [14]. In this study, we completed 2 activities (developing, validating, and testing ML models and studying model outputs with latent profile analysis [LPA]) and identified future activities toward optimization. The drivers behind this optimization initiative stem from low levels of patient engagement with dDPPs and other wellness-based mobile apps. We used the daily and weekly data sets provided by the dDPP vendor (inputs) to develop, validate, and test an ML model for each data set (first activity). On the basis of the performance metrics from the daily and weekly models, we identified the highest contributing feature for each model using Shapley values (first outputs). These features were fed into the LPA (second activity) to determine the number of participant usage profiles (second outputs). The goodness of fit derived from the LPA validated the phenotypes formed from the LPA (direct outcome). This integration of ML and statistical learning processes would inform how we identify digital engagement phenotypes for the dDPP study set (in the dashed red box) and, therefore, design content for a more personalized messaging platform (second direct outcome). Ultimately, the desired long-term outcomes of the profiling process are increased patient engagement with the dDPP and a reduction in clinical outcomes related to hemoglobin A_{1c} and weight (indirect outcomes). The process rests on the assumptions that the dDPP data accurately reflect digital behavioral patterns and that people from the vendor-provided data are representative of people in the study data set.

Figure 1. Logic diagram of the research methodology to integrate machine learning (ML) into participant profiling, including the input data sets; the methods applied to the data sets; and the intermediary, direct, and indirect outputs. dDPP: digital diabetes prevention program.



Participants

Study participants were users with prediabetes who enrolled in a commercial dDPP app (our dDPP research vendor), including nonpatient (“vendor”) users and institution-based patients (“study” participants of this dDPP intervention) [11]. Eligible participants are at least 18 years old, have a BMI of at least 25 kg/m² (22 kg/m² if self-identified as Asian), have a diagnosis of prediabetes (either by *International Classification of Diseases, Tenth Revision* code, problem list, or a hemoglobin A_{1c} level of 5.7%-6.4% in the last 12 months), and are deemed safe to engage in light physical exercise and weight loss by their primary care physician. For institutional study participants enrolled in the current clinical trial of this dDPP intervention, patients are excluded if they have a prior diagnosis of diabetes, have any end-stage illness with a prognosis within 6 months, are non-English speakers (as the dDPP program is currently only available in English), or are unable to send or receive SMS text messages [4]. Recruited patients were identified via electronic health record review and contacted through multichannel methods (eg, patient portal, email, in-clinic recruitment, and clinician referral).

The Data

Data Sourcing

Data for the evaluation were sourced from a commercial dDPP vendor and a patient cohort of an academic health center. We used 2 deidentified data sets (weekly and daily data) of eligible retail users for the initial training, validation, and testing of the ML models. These data sets aggregate and present user information on a weekly or daily basis and capture all features recorded by the dDPP app, including per user or patient: meals logged, steps logged, exercises logged, messages shared with the dDPP coach and other dDPP patients using the app, app log-ins, and the number of dDPP articles read. These activities were the same as those used for generating the adherence algorithm in our previous research. In addition to the

vendor-provided data sets, for a later testing phase, we use an existing data set of data collected from dDPP patients who are part of this dDPP study and exposed to the PAMS intervention.

Weekly dDPP Vendor Data Set

Data include detailed information about all the features collected for our dDPP app partners, such as meals logged, steps logged, exercises logged, messages shared with the dDPP coach and other dDPP patients using the app, app log-ins, and the number of dDPP articles read during each week. All users have more than 5 weeks of engagement records, and we used only 1 year’s worth of dDPP engagement data per user.

Weekly dDPP Institutional Study Data Set

The 2 data sets (weekly dDPP vendor data set and weekly dDPP study data set) have the same data structure. The same data fields are collected for commercial users and the dDPP patients, but the only difference is on the behavioral level because the patients’ data are potentially affected by the message intervention (PAMS). All data were used for the validation of the weekly ML model.

Daily dDPP Vendor Data Set

In addition to the activity records in the weekly data, we had access within the daily data set to calorie consumption data, meal logs, and color codes assigned to each food item as reported by the users. Users with less than 7 days of engagement records were excluded from the cohort, and we used only 1 year’s worth of dDPP engagement data per user.

Outcomes

First, we built binary classification ML models to predict whether a participant will engage in the next week or the next day with the dDPP based on their previous short- and long-term activity in the app. For the weekly model, we used the vendor data set to train and validate retrospectively to predict general activity (engagement in any activity). We prospectively validated the weekly model using the institutional study data

set. For the daily models, we predicted 5 outcomes: general activity, physical activities (steps and exercises recorded on the app), in-app lessons (article reading), social activities (group posts and coach messages in the app), and weigh-ins in the app. Second, we identified the variables from the daily overall activity model of the vendor's participants that provide the most predictive power for engagement. Third, we evaluated whether these predictive variables could generate profiles of a participant's behavior that can be targeted with motivational messaging.

Predictors

We built model predictors from users' demographic data and collected in-app activities. These activities include steps taken, exercises, meal logs, weigh-in records, in-app messaging and group activities, and in-app article reading. For the weekly data set, short-term activity profiles were built from the week before the evaluation week and up to 4 weeks before the evaluation week. Long-term activity profiles were summarized and constructed from the first week of program enrollment up to the evaluation week. Short-term activity profiles were built from the day before and within 7 days before the evaluation day for the daily data set. Similarly, long-term activity profiles were summarized and constructed from the first day of program enrollment up to the evaluation day. The day of the week and national holidays were also captured as predictors. In total, 43 predictors were used to build weekly models, and 49 predictors were used to build daily models.

Sample Size

The sample sizes for user weekly and daily data sets directly from the dDPP vendor were determined by the convenience of the dDPP vendor and assumed to be representative of the academic health center's study sample. The study sample size was determined by the number of participants already recruited and actively involved in the original dDPP study as of December 2021 [4].

Missing Data

Because this paper aims to predict participant engagement with the dDPP, missing data among in-app activities were treated as a participant not engaging in either overall activity (ie, no observations for a particular day or a week for any activity) or specific within-dDPP activities (eg, a participant not recording meals or reading any articles). Missing participant weight was logged as a participant not weighing themselves for the dDPP, and we ignored the magnitude of weight due to individual non-dDPP factors contributing to weight outcomes. No participant had a missing age due to age being a requirement for enrollment into the dDPP. Participants who did not record their ideal body weight at the beginning of dDPP engagement had this observation recorded as a 0, as the lack of goal recording for weight could have clinical implications (eg, weight is not the primary utilization goal for the participant, or the participant is not comfortable with setting a weight goal). No participant had a missing initial BMI recorded. One participant was missing gender identification, so their observations were removed from the data set.

Statistical Analysis Methods

Data Split

All data sets were split into a 70% training set, a 15% validation set, and a 15% test set based on users. Observations of any user only existed in 1 set to prevent potential data leak and unintended bias.

Gradient-Boosted Forest Algorithm

We use the gradient-boosted forest algorithm, a robust regression tree approach that includes multiple simple decision trees to iteratively refine the performance of the model by minimizing the difference between the expected and expert-labeled outcomes [15,16]. Forest-based algorithms provide 2 fundamental benefits. First, they allow for nonlinear interactions between covariates to impact the prediction of the dependent variable, as opposed to using a Least Absolute Shrinkage and Selection Operator (LASSO) or a ridge regression model. Second, forest-based algorithms do not require a priori function structure to define the relationship between the covariates and the outcome. For example, we do not need to theoretically assume whether a particular engagement type (eg, steps) interacts with another type (eg, exercise logging). We used gradient boosting to allow for prediction despite the sparsity of the data, as users may engage with one activity but not others on a given day or have no activity (ie, all observations as 0). The values defining engagement included binary predictors, large integers (eg, calories and steps), and values between 0 and 1 (eg, the portion of engagement throughout enrollment). These models aimed to identify that the sub-behaviors that create the most predictive power for engagement with the dDPP were trained with $\eta=0.1$ for 1000 rounds with early stopping.

Metrics

The area under the receiver operating characteristic curve (AUROC), the area under the precision-recall curve (AUPRC), and the Brier score statistics measured the performance of the model. To estimate the CIs of the evaluation metrics for the ML models, we performed bootstrapping with 200 iterations on the test set. In each iteration, a random sample of the test set, with replacement, was drawn with the same size as the original test set. The ML model was then evaluated on this bootstrapped sample, and the performance metrics mentioned above were recorded. The process was repeated for 200 iterations, resulting in a distribution of performance metrics from which the 95% CIs were calculated, providing a robust estimate of the performance and variability of the model. In addition, Shapley values were calculated to reflect the feature importance of each model.

Engagement Profiling

A person-centered approach to messaging can help motivate individuals to complete goal-oriented behaviors like engagement with a lifestyle management app [17]. This approach involves (1) tailoring delivery based on the person's behavior profile within the app and (2) focusing messaging on targetable behaviors to motivate users to complete small, manageable actions toward their goal (ie, the goal gradient hypothesis in decision-making) [18]. We performed an LPA on the participants in the daily data set to determine the subgroups of

participants' behaviors. LPA identifies latent clusters of individuals based on continuous variables [19]. The contributions of multiple variables (ie, the facets that explain the unobserved profile of a user) contribute to the outcome experienced by a user. We used the covariates with the highest global mean Shapley values from the gradient-boosted forest model for the LPA for 2 reasons. First, these variables offer the most explanatory power behind the probability of engagement with the dDPP, allowing us not to assume a priori the behaviors that contribute to the usage of the dDPP. Second, profiling users of a digital app such as this dDPP can be more complicated than traditional approaches to consumer profiling, given the interaction between a user's health and app engagement. To determine the minimum usage data after enrollment into a dDPP to start profiling participants, we conducted LPAs after 2 weeks and iteratively added days until 3 weeks of engagement. We used the profiles from the timestamp with the lowest Bayesian information criterion (BIC), the established goodness-of-fit metric for LPA. We used the *mclust* package in RStudio (version 2022.12.0+353; Posit Software, PBC) to run the LPAs [20].

Development Versus Validation

We validated the weekly model prospectively using the weekly dDPP study data set. Detailed information about this data set is under the subsection "Participants" [15,16].

Ethical Considerations

In this DPP research, ethical standards and the protection of human participants are emphasized. The study is committed to adhering to regulations outlined in 45 CFR Part 46, ensuring the rights and welfare of participants. The NYU Langone Health institutional review board (IRB) played a crucial role in reviewing and approving the research, informed consent forms, and recruitment materials before participant enrollment (i20-01548). The informed consent process is described as an ongoing dialogue, emphasizing clear communication,

comprehension, and the right to withdraw without adverse consequences. The consent forms, including verbal consent and a key information sheet, were submitted to the IRB for approval. Confidentiality measures are robust, complying with the Health Insurance Portability and Accountability Act (HIPAA), and a Certificate of Confidentiality from the National Institutes of Health was obtained. Data security is maintained through password protection, and research data are stored securely. The research emphasizes that stored data will only be used for this study, with no plans for future use in subsequent research. Overall, the research underscores the importance of ethical conduct, participant consent, and stringent confidentiality measures in the research process.

Moreover, the research underscores the importance of ethical conduct, rigorous IRB oversight, and robust confidentiality measures to safeguard the rights and well-being of study participants. Additionally, it highlights the meticulous documentation of the informed consent process and the secure handling of research data, ensuring compliance with regulations and promoting participant trust and privacy.

Results

Participants

Table 1 details the descriptive statistics for the 3 preprocessed data sets, including weekly and daily data for the dDPP user (dDPP vendor data sets) and the weekly data for the dDPP patients (dDPP study data). For the vendor-provided data sets, users engage with the app 54.2% (208,142/384,025) of the times in the weekly data compared with 38.9% (274,200/704,242) of the times in the daily data. The average engagement within individual activities is similar. "Steps taken" had the highest percentage of all activities in both data sets. For study data, the engagement percentage was higher (92.1%, 1253/1361), which could be attributed to the effects of PAMS messages.

Table 1. Descriptive statistics of users (N=12,262).

Characteristic	Weekly dDPP ^a vendor data (dDPP vendor users, n=10,053)	Weekly dDPP study data (dDPP study patients, n=50)	Daily dDPP vendor data (dDPP vendor users, n=2159)
Program length	38.2 weeks	27.22 weeks	326.2 days
Age (years), mean (SD)	47.6 (11.4)	N/A ^b	N/A
Sex, n (%)			
Male	1267 (12.6)	N/A	N/A
Female	8786 (87.4)	N/A	N/A
Engagement of any activity, n/N (%)	208,142/384,025 (54.2)	1253/1361 (92.1)	274,200/704,242 (38.9)
Engagement of steps taken, n/N (%)	208,142/384,025 (54.2)	1086/1361 (79.8)	244,823/704,242 (34.8)
Engagement of exercises, n/N (%)	77,957/384,025 (20.3)	349/1361 (25.6)	49,683/704,242 (7.1)
Engagement of meals logged, n/N (%)	137,865/384,025 (35.9)	924/1361 (67.9)	100,449/704,242 (14.3)
Engagement of weigh-ins, n/N (%)	137,481/384,025 (35.8)	523/1361 (38.4)	71,596/704,242 (10.2)
Engagement of article reading, n/N (%)	118,280/384,025 (30.8)	573/1361 (42.1)	79,272/704,242 (11.2)
Engagement of group posts, n/N (%)	24,578/384,025 (6.4)	100/1361 (7.3)	45,113/704,242 (6.4)

^adDPP: digital diabetes prevention program.

^bN/A: not applicable.

Weekly Model (for Any Activity) Development and Performance

We trained and tested the model to predict “any activity” (ie, the probability of the subsequent interaction with the dDPP based on whether the user interacted with any of the features of the dDPP app, such as exercise, meal, and weigh-ins) on the weekly dDPP vendor data set. The weekly model reported an AUROC of 0.97 (95% CI 0.97-0.97), an AUPRC of 0.98 (95% CI 0.98-0.98), and a Brier score of 0.061 (95% CI 0.060-0.063)

in the test set (Figure 2). Because we also aimed to identify how individual variables contribute to predictions by the model, we calculated the Shapley value, which is the average marginal contribution of a variable to a model across the different combinations of including the variable in the model (eg, nonlinear contributions and splitting a forest into different branches with the variable). The Shapley value method has become the preferred technique for feature attribution in ML models, thanks to its robust and reliable performance [21].

Figure 2. AUROC (left) and AUPRC (right) performance metrics of the “any activity” weekly model in the test set of the weekly vendor data set (58,210 engagement records). The calibration plot shows that the model is well calibrated. AUPRC: area under the precision-recall curve; AUROC: area under the receiver operating characteristic curve.

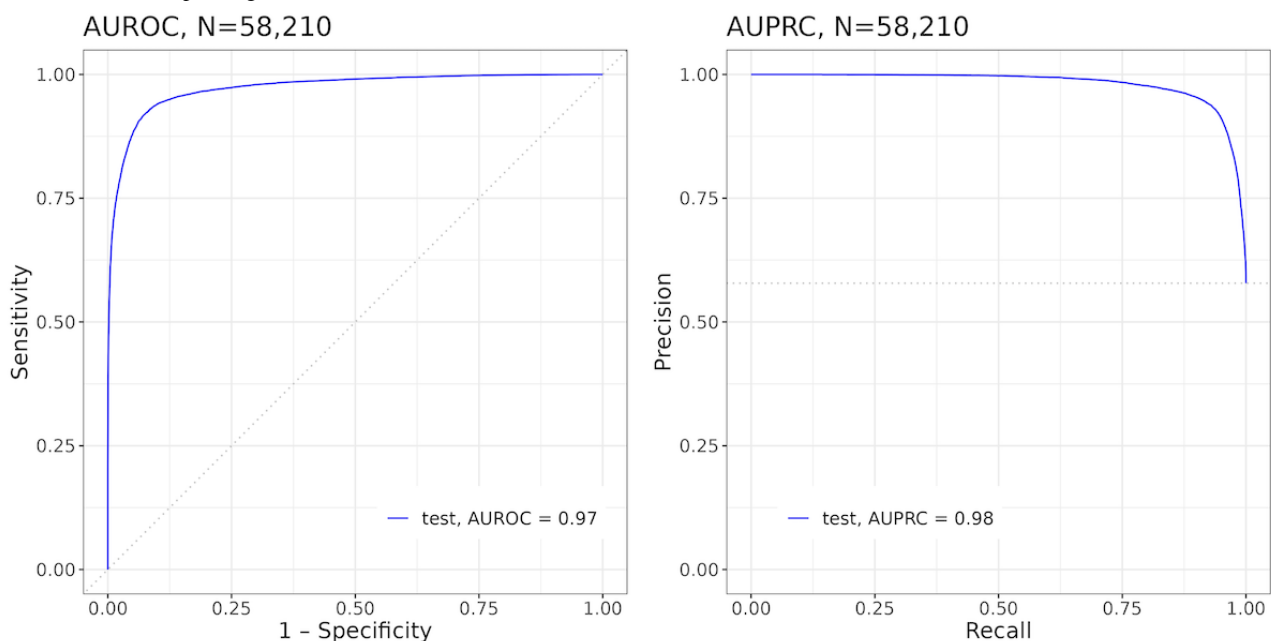
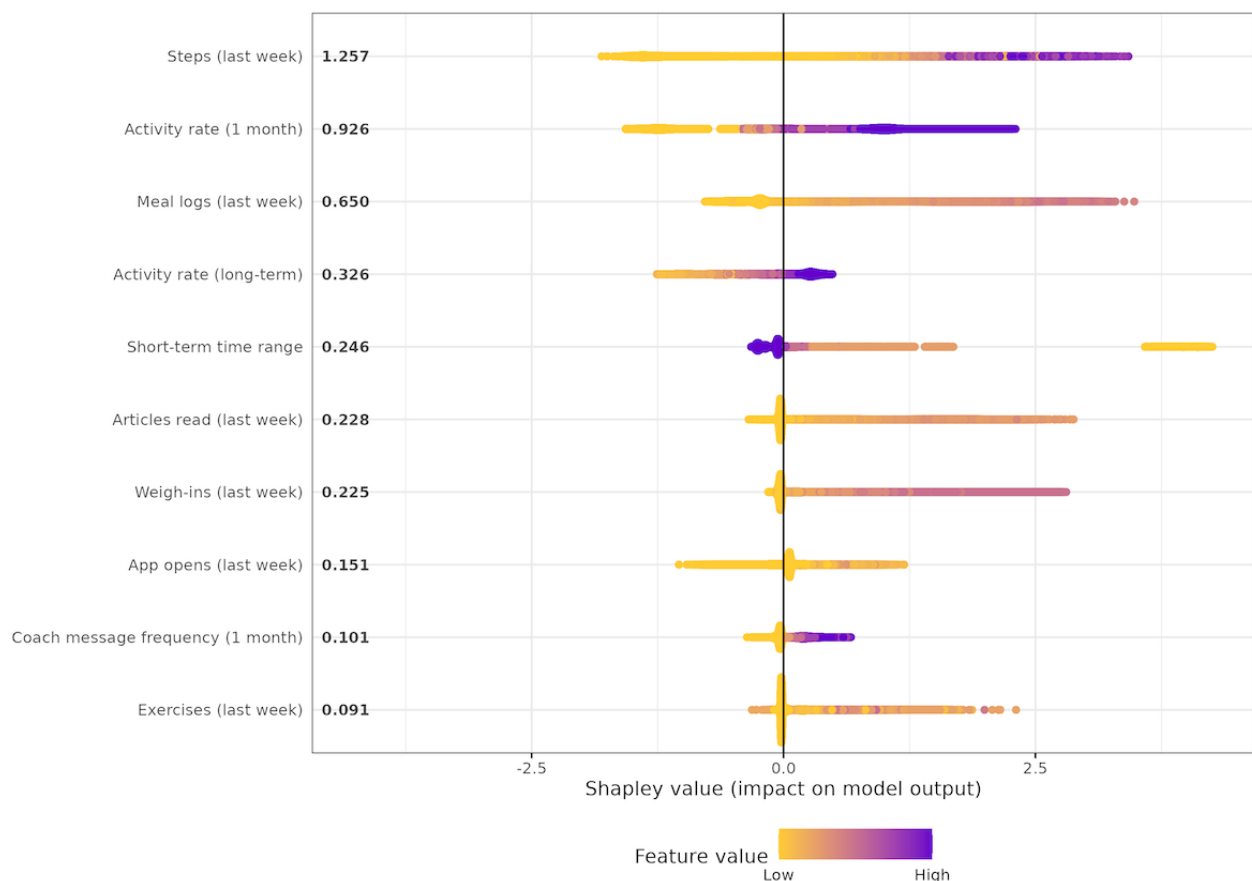


Figure 3 displays the distribution of the 10 covariates with the highest calculated global mean Shapley value (ie, which variables have the strongest predictive power, regardless of negative or positive impact, on the user's engagement with the dDPP). A higher magnitude of the Shapley value (ie, further from 0) indicates the strength of the variable in the model to predict a user's engagement with the dDPP. A positive Shapley value indicates that the user is more likely to engage with the dDPP because of the variable (ie, a positive predictor). A

negative Shapley value suggests that the patient is less likely to engage with the dDPP due to the variable (ie, a negative predictor). More purple values indicate a higher mean for the covariate of the individual (eg, a more purple "exercise frequency" dot indicates that the user logged for nonstep physical activity more than other users did). The covariates with the most contribution to model prediction were those of short-term behaviors.

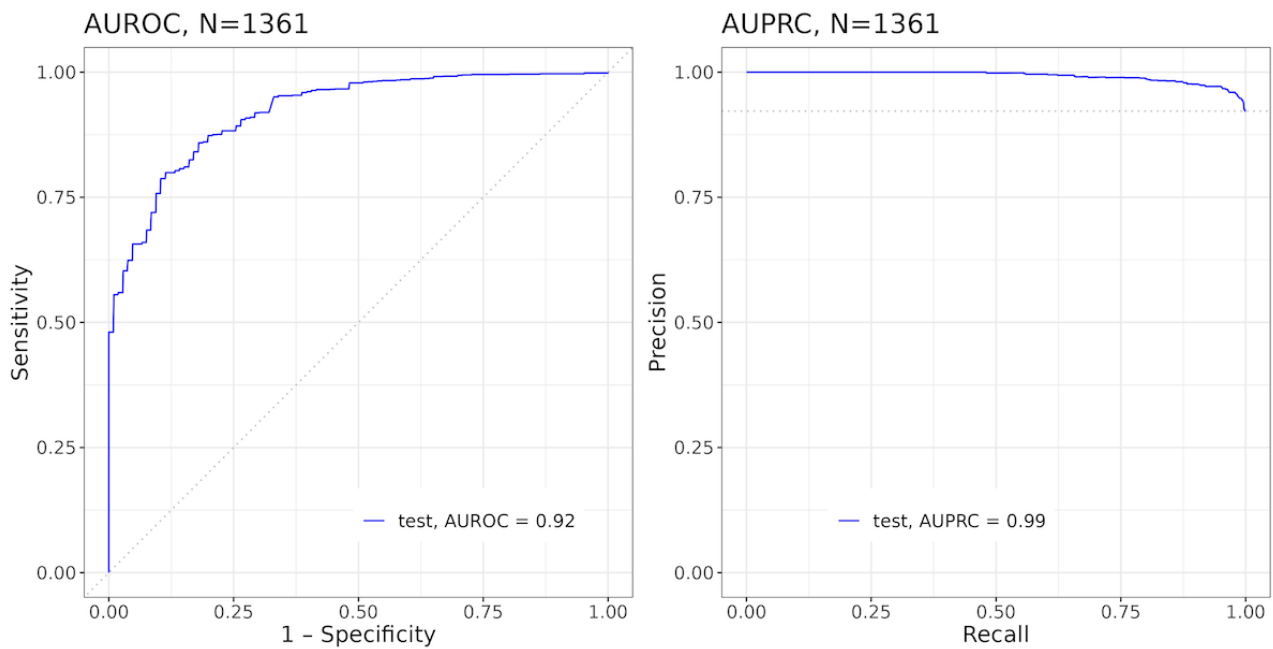
Figure 3. Shapley values of top 10 features in the "any activity weekly model." Each dot on the plot represents an engagement record and is colored according to the value of the corresponding feature from high (purple) to low (yellow). Features are ranked in descending order from top to bottom on the y-axis (ie, variables with the highest contribution to the model are on the top), with global mean Shapley values of each feature annotated next to them.



We tested our model using the weekly dDPP institutional study data set (prospective clinical data). The model achieved an AUROC of 0.92 (95% CI 0.89-0.94), an AUPRC of 0.99 (95% CI 0.99-0.99; Figure 4), and a Brier score of 0.072 (95% CI 0.063-0.081), suggesting high predictive power and operational potential for refining PAMS using this method. After analyzing the weekly dDPP study data set, we detected that this data set

would be imbalanced because the prediction of the subsequent week's activity would be based on whether a user engaged with any app activity, rather than a particular activity, within the dDPP, seen by the 92.1% engagement ratio, and the sample size was too low to yield unbiased testing results. Regardless of the limitation of the research data set, this analysis was proper in confirming the effectiveness of the weekly model.

Figure 4. AUROC (left) and AUPRC (right) performance metrics of any activity weekly model in the weekly study data set (1361 engagement records). AUPRC: area under the precision-recall curve; AUROC: area under the receiver operating characteristic curve.



Daily Model (for Any Activity) Development and Performance

We expanded a proportion of the weekly data set into a daily (more detailed) format and trained 5 new models. [Figure 5](#) illustrates the ML model fit in the test set of the daily data set. [Figure 6](#) displays the distribution of the covariates with the

strongest predictive power (ie, the highest global mean Shapley value). Like the weekly model, engagement with any activity in the dDPP in the last 7 days had the most predictive power (a global mean Shapley value of 2.638). However, in contrast to the weekly model, features associated with long-term activity also had strong predictive power in the model.

Figure 5. AUROC (left) and AUPRC (right) performance metrics of the “any activity” daily model in the test set of the daily vendor data set (106,950 engagement records). AUPRC: area under the precision-recall curve; AUROC: area under the receiver operating characteristic curve.

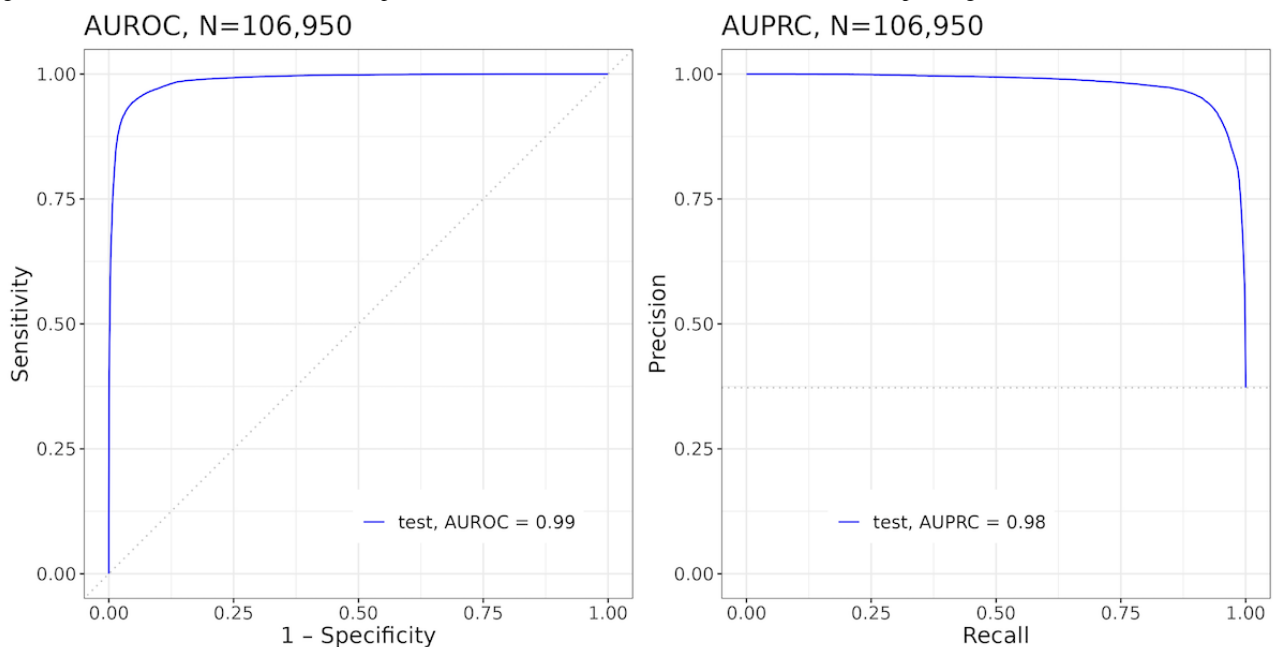
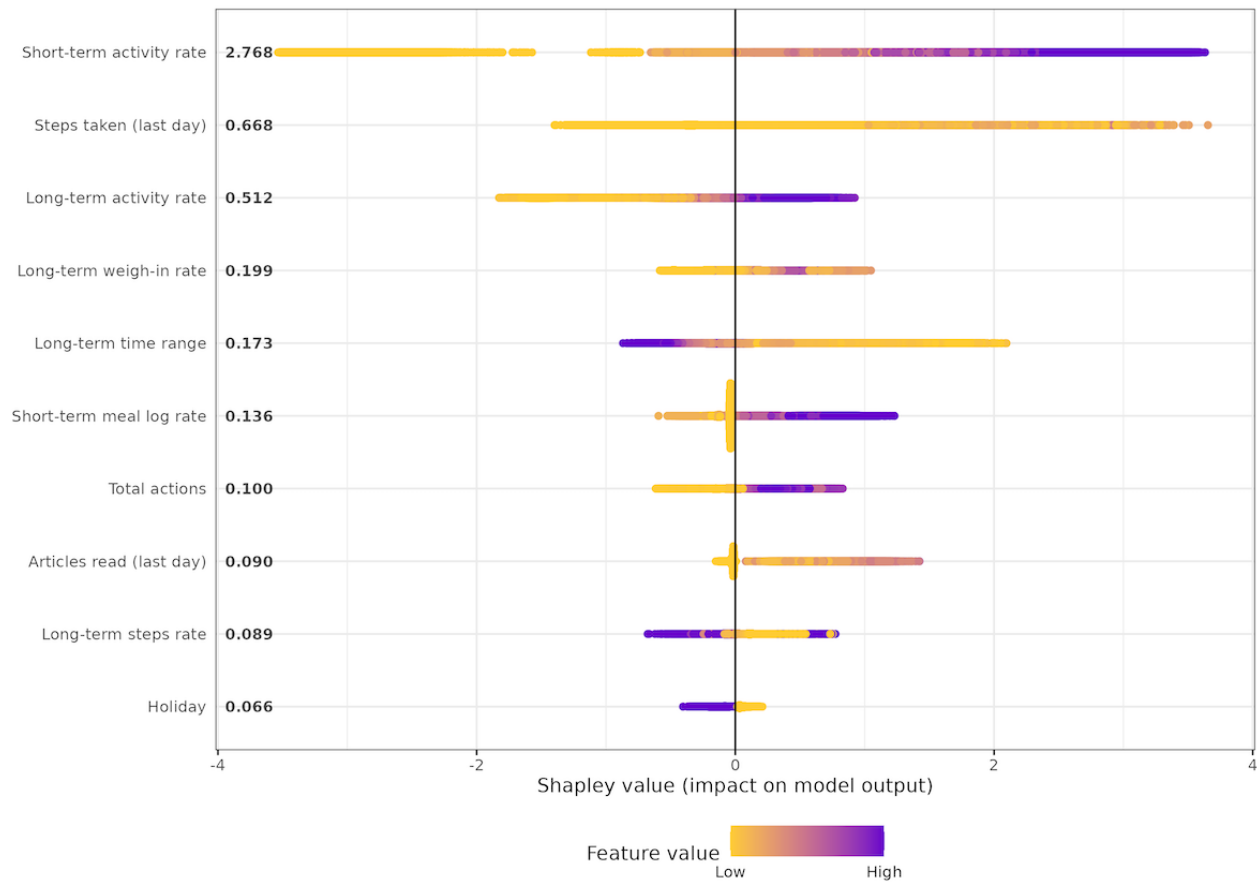


Figure 6. Shapley values of top 10 features in the “any activity” daily model. Each dot on the plot represents an engagement record and is colored according to the value of the corresponding feature from high (purple) to low (yellow). Features are ranked in descending order from top to bottom on the y-axis. Average Shapley values of each feature are annotated next to them on the y-axis.



Although the daily model for “any activity” returned a high AUROC and AUPRC, we aimed to generate predictions on each specific activity to inform our user profiling (digital engagement phenotypes) and consequently elevate the message personalization. Therefore, we developed 4 ML models, focusing on daily engagement with each key type of activity for a dDPP (physical activity, lessons, social activity, and weigh-ins). Table 2 displays the model fits for each of these

“submodels.” For each activity, the model indicates highly predictive behavioral patterns among users. The “physical activity” and “social activity” daily models had higher AUROC performance with slightly lower AUPRC than the other daily models. All daily models show higher levels of calibration (a highest Brier score of 0.051) than the weekly model (a Brier score of 0.061).

Table 2. Performance metrics of each daily activity model in the test set.

Model fit metrics	Any app activity	Physical activity (exercises and steps)	Lessons (article reading)	Social activity (group posts and coach messages)	Weigh-ins
AUROC ^a (95% CI)	0.99 (0.99-0.99)	0.98 (0.98-0.98)	0.99 (0.99-0.99)	0.98 (0.98-0.98)	0.94 (0.94-0.94)
AUPRC ^b (95% CI)	0.98 (0.98-0.98)	0.74 (0.72-0.75)	0.91 (0.91-0.92)	0.74 (0.73-0.75)	0.65 (0.63-0.66)
Brier score (95% CI)	0.037 (0.036-0.038)	0.025 (0.025-0.026)	0.027 (0.026-0.028)	0.02 (0.023-0.024)	0.051 (0.050-0.052)

^aAUROC: area under the receiver operating characteristic curve.

^bAUPRC: area under the precision-recall curve.

Engagement Profiling Development and Performance

We profiled participants with their daily engagement data using LPA after 2 weeks of dDPP enrollment. To determine the optimal time to start profiling participants, we iteratively added 1 day of engagement and created profiles until 3 weeks after

their enrollment in the dDPP. After 2 weeks of daily engagement data, profiling participants had the strongest LPA model fit (BIC=-3222.46), followed by the model fit from profiling with 3 weeks of data (BIC=-2903.19). The LPA model fits for 15 to 20 days of engagement were significantly worse (ie, higher

BIC values) and, therefore, are not reported. The best-performing LPA model was ellipsoidal (there is some correlation between variables), had equal volume (the variances are equal across identified profiles), had variable distributions

between profiles (ie, the number of people per profile vary), and consisted of 6 profiles. Table 3 reports the mean engagement for each variable within and across the profiles of participants.

Table 3. Mean engagement by profile and across profiles for key engagement variables.

Key engagement variables	Subbehavior variable mean (SE)						Mean engagement across profiles (SD)
	Profile 1 (n=16)	Profile 2 (n=91)	Profile 3 (n=107)	Profile 4 (n=20)	Profile 5 (n=82)	Profile 6 (n=8)	
Any activity rate (last 7 days)	0.969 (0.085)	0.992 (0.045)	1.000 (0)	0.747 (0.243)	0.115 (0.236)	0.814 (0.222)	0.752 (0.401)
Long-term activity rate	0.942 (0.105)	0.998 (0.013)	0.979 (0.049)	0.605 (0.262)	0.365 (0.292)	0.698 (0.244)	0.797 (0.318)
Steps taken rate (last 7 days)	2572 (4661)	3909 (3968)	3646 (3461)	1378 (1293)	0 (0)	1622 (2447)	2555 (3495)
Long-term weigh-in rate	0.507 (0.310)	0.139 (0.139)	0.265 (0.208)	0.0362 (0.043)	0.0471 (0.131)	0.241 (0.175)	0.172 (0.208)
Recent meal rate	0.906 (0.256)	0.397 (0.416)	0.690 (0.399)	0.0252 (0.112)	0.00305 (0.028)	0.00305 (0.297)	0.392 (0.444)
Long-term step rate	0.438 (0.345)	0.998 (0.013)	0.956 (0.068)	0.566 (0.292)	0.285 (0.295)	0.609 (0.303)	0.740 (0.359)
Long-term meal log rate	0.856 (0.249)	0.463 (0.339)	0.669 (0.349)	0.0543 (0.127)	0.0951 (0.145)	0.116 (0.076)	0.425 (0.386)
Article reading rate (last 7 days)	2.01 (1.549)	0.278 (0.704)	2.85 (1.644)	3.021 (0.923)	0 (0)	0.372 (1.061)	1.160 (1.689)

The LPA identified attrition (users in profile 5 who showed consistently low engagement across variables) and behaviors that show points of continued engagement for users. Users in profile 6, for example, had a close-to-average engagement with the dDPP from weigh-ins with the app and logging steps, which are behaviors that require one-time interactions with the dDPP, given Bluetooth connections between smart devices and the dDPP. In contrast, users in profile 3 were highly engaged, as they consistently engaged more than the average user. Messaging to users in profile 3 should, therefore, differ from messaging to users in profile 5, given the differences in their efforts toward the dDPP. Users in profile 4 had a lower-than-average engagement with the dDPP but showed the highest engagement with the learning materials across all users. Clusters 1 and 2 showed similarly high short- and long-term engagements but differed in engagement with the dDPP. Users in profile 1 read more educational materials provided in the dDPP, whereas users in profile 2 were more consistent in taking steps.

Discussion

Summary

The literature suggests the app of different ML algorithms to predict digital and traditional medication adherence and diverse intervention outcomes. Positive results of these studies support and validate the feasibility of applying ML methods to predict user engagement in digital health apps such as a dDPP to improve patient adherence to digital therapeutics and, consequently, health outcomes. In concordance with the literature, we applied the most suitable algorithm for our data

set (gradient-boosted forest), yielded highly accurate results for predicting digital adherence, and identified variables with the strongest contribution to our outcome to understand digital behaviors [22-26]. This paper described 2 ML models developed using weekly and daily dDPP engagement data. First, using the weekly dDPP vendor data set, we developed a weekly ML model, which was validated using the collected data from this dDPP study. On the basis of past activity patterns, the model yielded high precision and recall and accurately predicted patient engagement for the next week. However, a model trained with weekly patient data can only predict weekly engagement, limiting our ability to gain detailed insight into a patient's behavior. Because an ideal model should be robust to different dynamics in patients' engagement data, we then developed a daily ML model using the daily dDPP vendor data set, which incorporates additional attributes, including the type of meals logged per day and calories. The daily model also yielded high precision and recall values. This finding supports using such models to anticipate behavior, focusing on identifying low engagement to intervene before attrition.

In addition to calculating precision and recall for our models, we calculated the Shapley values for both types of models (weekly and daily) to further analyze and identify which variables contribute the most to overall prediction. Results from the Shapley values revealed that short-term frequency of activity engagement was the most informative feature in the daily and weekly data analyses, meaning that users were more likely to form and stick to short-term behavioral patterns than long-term patterns in the dDPP. This finding is consistent with a previous study on predicting exercise and steps [27]. Because of user propensity to engage in short-term behaviors, we considered

the daily model for individual activities best suited to develop engagement profiles. Using variables with high Shapley values from the daily model, we successfully created distinct digital engagement phenotypes of dDPP users. This allows for further research into developing infrastructure for tailored messaging to increase and maintain engagement with active users and intervene against attrition for inactive users. Specifically, identifying high engagement, minimal engagement, and attrition with early dDPP use lends itself to determining individuals facing barriers to dDPP engagement and improving dDPP implementation. Identifying strengths and weaknesses within behavior phenotypes through our profiling methods can also inform what specific behaviors (ie, low-engagement behaviors) need to be targeted in messaging for a user's success in using the dDPP.

Contributions and Implications

By leveraging digital behavioral usage data, we showed that we can successfully create digital engagement phenotypes, allowing for the future tailoring of digital health interventions based on patient needs. The methods used can extend beyond the prevention of metabolic disease, as an ML model incorporating behavioral usage variables can characterize prevention, maintenance, and wellness in other domains such as mental health, treatment adherence, and addiction prevention.

Limitations

The weekly data sets posed limitations to maximizing patient engagement through integrating ML into PAMS. A model trained using weekly data is limited to predict weekly dDPP engagement (limited scope of dDPP engagement). The weekly ML model did not provide enough granularity to be robust to different dynamics of app engagement (eg, a sudden drop in engagement in 1 week due to vacation or a suddenly busy day

where the user does not log information). The high sensitivity in a weekly engagement model to unexpected changes in usage could, therefore, negatively impact the type of messaging and timely motivation delivered to the patient. Consequently, we shifted the prediction cycle for engagement by moving from a model based on weekly behavior to one based on daily behavior.

Data showed that the short-term frequency of various activities was the most informative feature, but the results could mean that our model is vulnerable to short-term disruption of user behavioral patterns. Consequently, although the weekly data-based and daily data-based models were sufficient to prove the feasibility of using ML approaches for predicting patient engagement, further development is needed to refine these models and include extra patient information. Improvements include (1) understanding potential errors in the model and data sets (eg, data set size; using vendor data sets is an imperfect representation of other dDPP interventions) and (2) reviewing initial hypotheses about the data set and the choice of algorithms. To build the refined model, we would benefit from more detailed data. In this case, we would need to replan attributes and test other ML algorithms to perform further model improvements.

Future Directions

With feasibility established, the next steps include creating user engagement phenotypes linked to personalized messaging interventions using behavior-based approaches to best motivate users to engage with the dDPP. We will also need to engineer the forest model and profile analysis to evolve as users change their engagement throughout participating in the dDPP so that messaging remains personalized to meet the users' needs. Ultimately, this study demonstrated the potential value of ML and digital phenotyping to enhance the ability of digital behavior change interventions to predict engagement and personalize the interventions to maximize clinical impact.

Acknowledgments

Noom, Inc provided data from their commercial digital diabetes prevention program (dDPP) users, which was used as a baseline to train our machine learning model and obtain preliminary results. We thank Nina Singh for her feedback on the manuscript. This work was supported by the National Institute of Diabetes and Digestive and Kidney Diseases (grant 1R18DK118545-01A1; Principal Investigator: DMM). RVNV is funded by HRSA Ruth L Kirschstein National Research Service Award (ID T32HP22238).

Authors' Contributions

DVR, JC, and RVNV made substantial contributions to the conception or design of the work, as well as the acquisition, analysis, or interpretation of data for the work. They also contributed to drafting the work or revising it critically for important intellectual content. JC and RVNV contributed to the development of machine learning models and data analysis. All authors agreed to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved. All authors gave their final approval of the version to be published.

Conflicts of Interest

None declared.

References

1. Laranjo L, Dunn AG, Tong HL, Kocaballi AB, Chen J, Bashir R, et al. Conversational agents in healthcare: a systematic review. *J Am Med Inform Assoc* 2018;25(9):1248-1258 [FREE Full text] [doi: [10.1093/jamia/ocy072](https://doi.org/10.1093/jamia/ocy072)] [Medline: [30010941](https://pubmed.ncbi.nlm.nih.gov/30010941/)]

2. Oh YJ, Zhang J, Fang ML, Fukuoka Y. A systematic review of artificial intelligence chatbots for promoting physical activity, healthy diet, and weight loss. *Int J Behav Nutr Phys Act* 2021;18(1):160 [FREE Full text] [doi: [10.1186/s12966-021-01224-6](https://doi.org/10.1186/s12966-021-01224-6)] [Medline: [34895247](https://pubmed.ncbi.nlm.nih.gov/34895247/)]
3. Friedman R, Sedoc J, Gretz S, Toledo A, Weeks R, Bar-Zeev N, et al. VIRATrustData: a trust-annotated corpus of human-chatbot conversations about COVID-19 vaccines. ArXiv. Preprint posted online on May 24, 2022 2022 [FREE Full text]
4. Rodriguez DV, Lawrence K, Luu S, Yu JL, Feldthouse DM, Gonzalez J, et al. Development of a computer-aided text message platform for user engagement with a digital Diabetes Prevention Program: a case study. *J Am Med Inform Assoc* 2021;29(1):155-162 [FREE Full text] [doi: [10.1093/jamia/ocab206](https://doi.org/10.1093/jamia/ocab206)] [Medline: [34664647](https://pubmed.ncbi.nlm.nih.gov/34664647/)]
5. Dimitrov DV. Medical internet of things and big data in healthcare. *Healthc Inform Res* 2016;22(3):156-163 [FREE Full text] [doi: [10.4258/hir.2016.22.3.156](https://doi.org/10.4258/hir.2016.22.3.156)] [Medline: [27525156](https://pubmed.ncbi.nlm.nih.gov/27525156/)]
6. Horrigan JB, Duggan M. Home broadband 2015. Pew Research Center. 2015. URL: <https://www.pewresearch.org/internet/2015/12/21/home-broadband-2015/> [accessed 2016-10-14]
7. Kohl LF, Crutzen R, de Vries NK. Online prevention aimed at lifestyle behaviors: a systematic review of reviews. *J Med Internet Res* 2013;15(7):e146 [FREE Full text] [doi: [10.2196/jmir.2665](https://doi.org/10.2196/jmir.2665)] [Medline: [23859884](https://pubmed.ncbi.nlm.nih.gov/23859884/)]
8. Levey NN. Medical professionalism and the future of public trust in physicians. *JAMA* 2015;313(18):1827-1828. [doi: [10.1001/jama.2015.4172](https://doi.org/10.1001/jama.2015.4172)] [Medline: [25965228](https://pubmed.ncbi.nlm.nih.gov/25965228/)]
9. Alkhalidi G, Hamilton FL, Lau R, Webster R, Michie S, Murray E. The effectiveness of technology-based strategies to promote engagement with digital interventions: a systematic review protocol. *JMIR Res Protoc* 2015;4(2):e47 [FREE Full text] [doi: [10.2196/resprot.3990](https://doi.org/10.2196/resprot.3990)] [Medline: [25921274](https://pubmed.ncbi.nlm.nih.gov/25921274/)]
10. McTigue KM, Bhargava T, Bryce CL, Conroy M, Fischer GS, Hess R, et al. Patient perspectives on the integration of an intensive online behavioral weight loss intervention into primary care. *Patient Educ Couns* 2011;83(2):261-264. [doi: [10.1016/j.pec.2010.05.009](https://doi.org/10.1016/j.pec.2010.05.009)] [Medline: [21459256](https://pubmed.ncbi.nlm.nih.gov/21459256/)]
11. Lawrence K, Rodriguez DV, Feldthouse DM, Shelley D, Yu JL, Belli HM, et al. Effectiveness of an integrated engagement support system to facilitate patient use of digital diabetes prevention programs: protocol for a randomized controlled trial. *JMIR Res Protoc* 2021;10(2):e26750 [FREE Full text] [doi: [10.2196/26750](https://doi.org/10.2196/26750)] [Medline: [33560240](https://pubmed.ncbi.nlm.nih.gov/33560240/)]
12. Rodriguez DV, Lawrence K, Luu S, Chirn B, Gonzalez J, Mann D. PAMS—a personalized automatic messaging system for user engagement with a digital diabetes prevention program. : IEEE; 2022 Presented at: 2022 IEEE 10th International Conference on Healthcare Informatics (ICHI); June 11-14, 2022; Rochester, MN p. 297-308. [doi: [10.1109/ichi54592.2022.00051](https://doi.org/10.1109/ichi54592.2022.00051)]
13. Sherwin J, Lawrence K, Gragnano V, Testa PA. Scaling virtual health at the epicentre of coronavirus disease 2019: a case study from NYU Langone Health. *J Telemed Telecare* 2022;28(3):224-229 [FREE Full text] [doi: [10.1177/1357633X20941395](https://doi.org/10.1177/1357633X20941395)] [Medline: [32686555](https://pubmed.ncbi.nlm.nih.gov/32686555/)]
14. New Zealand SuPERU (Issuing body). Making Sense of Evaluation: A Handbook for Everyone: Using Evidence for Impact. Wellington: SuPERU; 2017.
15. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. New York, NY: Association for Computing Machinery; 2016 Presented at: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; August 13-17, 2016; San Francisco, CA p. 785-794. [doi: [10.1145/2939672.2939785](https://doi.org/10.1145/2939672.2939785)]
16. Lu Z, Sim JA, Wang JX, Forrest CB, Krull KR, Srivastava D, et al. Natural language processing and machine learning methods to characterize unstructured patient-reported outcomes: validation study. *J Med Internet Res* 2021;23(11):e26777 [FREE Full text] [doi: [10.2196/26777](https://doi.org/10.2196/26777)] [Medline: [34730546](https://pubmed.ncbi.nlm.nih.gov/34730546/)]
17. Veazie PJ, Cai S. A connection between medication adherence, patient sense of uniqueness, and the personalization of information. *Med Hypotheses* 2007;68(2):335-342. [doi: [10.1016/j.mehy.2006.04.077](https://doi.org/10.1016/j.mehy.2006.04.077)] [Medline: [17008025](https://pubmed.ncbi.nlm.nih.gov/17008025/)]
18. Sorrentino RM, Short JAC, Raynor JO. Uncertainty orientation: implications for affective and cognitive views of achievement behavior. *J Pers Soc Psychol* 1984;46(1):189-206. [doi: [10.1037//0022-3514.46.1.189](https://doi.org/10.1037//0022-3514.46.1.189)]
19. Wardenaar K. Latent profile analysis in R: a tutorial and comparison to Mplus. PsyArXiv. Preprint posted online on April 9, 2021 2021 [FREE Full text] [doi: [10.31234/osf.io/wzftf](https://doi.org/10.31234/osf.io/wzftf)]
20. Fraley C, Raftery AE, Scrucca L, Murphy TB, Fop M. mclust: Gaussian Mixture Modelling for Model-Based Clustering, Classification, and Density Estimation. R package version 5.2. 2016. URL: <https://CRAN.R-project.org/package=mclust>
21. Merrick L, Taly A. The explanation game: explaining machine learning models using shapley values. In: Holzinger A, Kieseberg P, Tjoa A, Weippl E, editors. Machine Learning and Knowledge Extraction: 4th IFIP TC 5, TC 12, WG 8.4, WG 8.9, WG 12.9 International Cross-Domain Conference, CD-MAKE 2020, Dublin, Ireland, August 25-28, 2020, Proceedings. Cham: Springer; 2020:17-38.
22. Bohlmann A, Mostafa J, Kumar M. Machine learning and medication adherence: scoping review. *JMIRx Med* 2021;2(4):e26993 [FREE Full text] [doi: [10.2196/26993](https://doi.org/10.2196/26993)] [Medline: [37725549](https://pubmed.ncbi.nlm.nih.gov/37725549/)]
23. Wang L, Fan R, Zhang C, Hong L, Zhang T, Chen Y, et al. Applying machine learning models to predict medication nonadherence in Crohn's disease maintenance therapy. *Patient Prefer Adherence* 2020;14:917-926 [FREE Full text] [doi: [10.2147/PPA.S253732](https://doi.org/10.2147/PPA.S253732)] [Medline: [32581518](https://pubmed.ncbi.nlm.nih.gov/32581518/)]

24. Hornstein S, Forman-Hoffman V, Nazander A, Ranta K, Hilbert K. Predicting therapy outcome in a digital mental health intervention for depression and anxiety: a machine learning approach. *Digit Health* 2021;7:20552076211060659 [FREE Full text] [doi: [10.1177/20552076211060659](https://doi.org/10.1177/20552076211060659)] [Medline: [34868624](https://pubmed.ncbi.nlm.nih.gov/34868624/)]
25. Lu HY, Ding X, Hirst JE, Yang Y, Yang J, Mackillop L, et al. Digital health and machine learning technologies for blood glucose monitoring and management of gestational diabetes. *IEEE Rev Biomed Eng* 2024;17:98-117 [FREE Full text] [doi: [10.1109/RBME.2023.3242261](https://doi.org/10.1109/RBME.2023.3242261)] [Medline: [37022834](https://pubmed.ncbi.nlm.nih.gov/37022834/)]
26. Javaid A, Zghyer F, Kim C, Spaulding EM, Isakadze N, Ding J, et al. Medicine 2032: the future of cardiovascular disease prevention with machine learning and digital health technology. *Am J Prev Cardiol* 2022;12:100379 [FREE Full text] [doi: [10.1016/j.ajpc.2022.100379](https://doi.org/10.1016/j.ajpc.2022.100379)] [Medline: [36090536](https://pubmed.ncbi.nlm.nih.gov/36090536/)]
27. Zhou M, Fukuoka Y, Goldberg K, Vittinghoff E, Aswani A. Applying machine learning to predict future adherence to physical activity programs. *BMC Med Inform Decis Mak* 2019;19(1):169 [FREE Full text] [doi: [10.1186/s12911-019-0890-0](https://doi.org/10.1186/s12911-019-0890-0)] [Medline: [31438926](https://pubmed.ncbi.nlm.nih.gov/31438926/)]

Abbreviations

AI: artificial intelligence
AUPRC: area under the precision-recall curve
AUROC: area under the receiver operating characteristic curve
BIC: Bayesian information criterion
dDPP: digital diabetes prevention program
DPP: diabetes prevention program
HIPAA: Health Insurance Portability and Accountability Act
IRB: institutional review board
LASSO: Least Absolute Shrinkage and Selection Operator
LPA: latent profile analysis
ML: machine learning
PAMS: personalized automatic messaging system

Edited by C Xiao; submitted 08.03.23; peer-reviewed by D Whitehead, J Sussman; comments to author 05.06.23; revised version received 25.07.23; accepted 03.01.24; published 01.03.24.

Please cite as:

Rodriguez DV, Chen J, Viswanadham RVN, Lawrence K, Mann D

Leveraging Machine Learning to Develop Digital Engagement Phenotypes of Users in a Digital Diabetes Prevention Program: Evaluation Study

JMIR AI 2024;3:e47122

URL: <https://ai.jmir.org/2024/1/e47122>

doi: [10.2196/47122](https://doi.org/10.2196/47122)

PMID: [38875579](https://pubmed.ncbi.nlm.nih.gov/38875579/)

©Danissa V Rodriguez, Ji Chen, Ratnalekha V N Viswanadham, Katharine Lawrence, Devin Mann. Originally published in *JMIR AI* (<https://ai.jmir.org/>), 01.03.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in *JMIR AI*, is properly cited. The complete bibliographic information, a link to the original publication on <https://www.ai.jmir.org/>, as well as this copyright and license information must be included.

Original Paper

Behavioral Nudging With Generative AI for Content Development in SMS Health Care Interventions: Case Study

Rachel M Harrison¹, BA; Ekaterina Lapteva², PhD; Anton Bibin³, PhD

¹GenAI Lab, Ophiuchus LLC, Dover, DE, United States

²Institute of Psychology, Russian Academy of Sciences, Moscow, Russian Federation

³Skoltech AI (Centers for Research, Education, and Innovation), Skolkovo Institute of Science and Technology, Moscow, Russian Federation

Corresponding Author:

Rachel M Harrison, BA

GenAI Lab

Ophiuchus LLC

1111B S Governors Ave

STE 7359

Dover, DE, 19904

United States

Phone: 1 302 526 0926

Email: rae@ophiuchus.ai

Abstract

Background: Brief message interventions have demonstrated immense promise in health care, yet the development of these messages has suffered from a dearth of transparency and a scarcity of publicly accessible data sets. Moreover, the researcher-driven content creation process has raised resource allocation issues, necessitating a more efficient and transparent approach to content development.

Objective: This research sets out to address the challenges of content development for SMS interventions by showcasing the use of generative artificial intelligence (AI) as a tool for content creation, transparently explaining the prompt design and content generation process, and providing the largest publicly available data set of brief messages and source code for future replication of our process.

Methods: Leveraging the pretrained large language model GPT-3.5 (OpenAI), we generate a collection of messages in the context of medication adherence for individuals with type 2 diabetes using evidence-derived behavior change techniques identified in a prior systematic review. We create an attributed prompt designed to adhere to content (readability and tone) and SMS (character count and encoder type) standards while encouraging message variability to reflect differences in behavior change techniques.

Results: We deliver the most extensive repository of brief messages for a singular health care intervention and the first library of messages crafted with generative AI. In total, our method yields a data set comprising 1150 messages, with 89.91% (n=1034) meeting character length requirements and 80.7% (n=928) meeting readability requirements. Furthermore, our analysis reveals that all messages exhibit diversity comparable to an existing publicly available data set created under the same theoretical framework for a similar setting.

Conclusions: This research provides a novel approach to content creation for health care interventions using state-of-the-art generative AI tools. Future research is needed to assess the generated content for ethical, safety, and research standards, as well as to determine whether the intervention is successful in improving the target behaviors.

(JMIR AI 2024;3:e52974) doi:[10.2196/52974](https://doi.org/10.2196/52974)

KEYWORDS

generative artificial intelligence; generative AI; prompt engineering; large language models; GPT; content design; brief message interventions; mHealth; behavior change techniques; medication adherence; type 2 diabetes

Introduction

Overview

Health care interventions involving written communication play a pivotal role in disseminating critical information to patients and promoting positive health outcomes. However, the process of crafting effective health care content has historically been labor-intensive, time-consuming, and often lacks the necessary uniformity and transparency required for rigorous research and development.

We propose the application of generative artificial intelligence (AI) technologies to address the pressing need for efficient and transparent content creation in health care interventions. In particular, we focus on harnessing the capabilities of pretrained large language models (LLMs), which are sophisticated AI systems designed to understand and generate human-like text (refer to subsection Generative AI With LLMs). By using these rapidly growing technologies, we aim to assist researchers in the content creation process, making it more accessible, systematic, and adaptable. As a tangible example, we introduce the first publicly available data set of AI-generated brief messages tailored for individuals with type 2 diabetes, specifically targeting medication adherence, a critical aspect of diabetes management. Notably, our data set of 1150 messages also stands as the current largest data set of health care intervention messages publicly available. Furthermore, we make our source code replicable and accessible to the research community while providing a comprehensive breakdown of our design process. In doing so, we seek to use generative AI to pave the way for a new era of health care intervention content development, one characterized by transparency, efficiency, and scientific rigor. Our main contributions are as follows:

1. Present a generative AI approach to content creation in brief message health care interventions
2. Illustrate the process of prompt engineering for content design within a particular theoretical framework
3. Provide the first publicly available data set of AI-generated intervention messages and release the source code as a resource for future research.

Mobile Health Interventions

In the ever-growing landscape of health care, effective communication is essential to enhancing preventive measures and developing intervention strategies that improve public health outcomes. Each year, a great number of new intervention studies are added to the health care literature [1,2]. However, with the growth in the quantity of interventions often comes an increase in their technical complexity, especially in the area of mobile health (mHealth). Many of these interventions are delivered through proprietary apps or other nonstandardized platforms, which complicates their integration into future programs and often causes their results to be obfuscated by the unique specifics of their deployment. This not only makes them challenging to apply elsewhere but also ensures that their development is both time-consuming and resource-intensive [3]. In addition, while research into mHealth has boomed in the last decade, studies suggest that the overall success rate of most mHealth interventions is not exactly clear, despite the strong interest and

obvious potential such interventions have [4,5]. Ensuring that these interventions are feasible, effective, and sustainable, is vital for preventing unnecessary research waste.

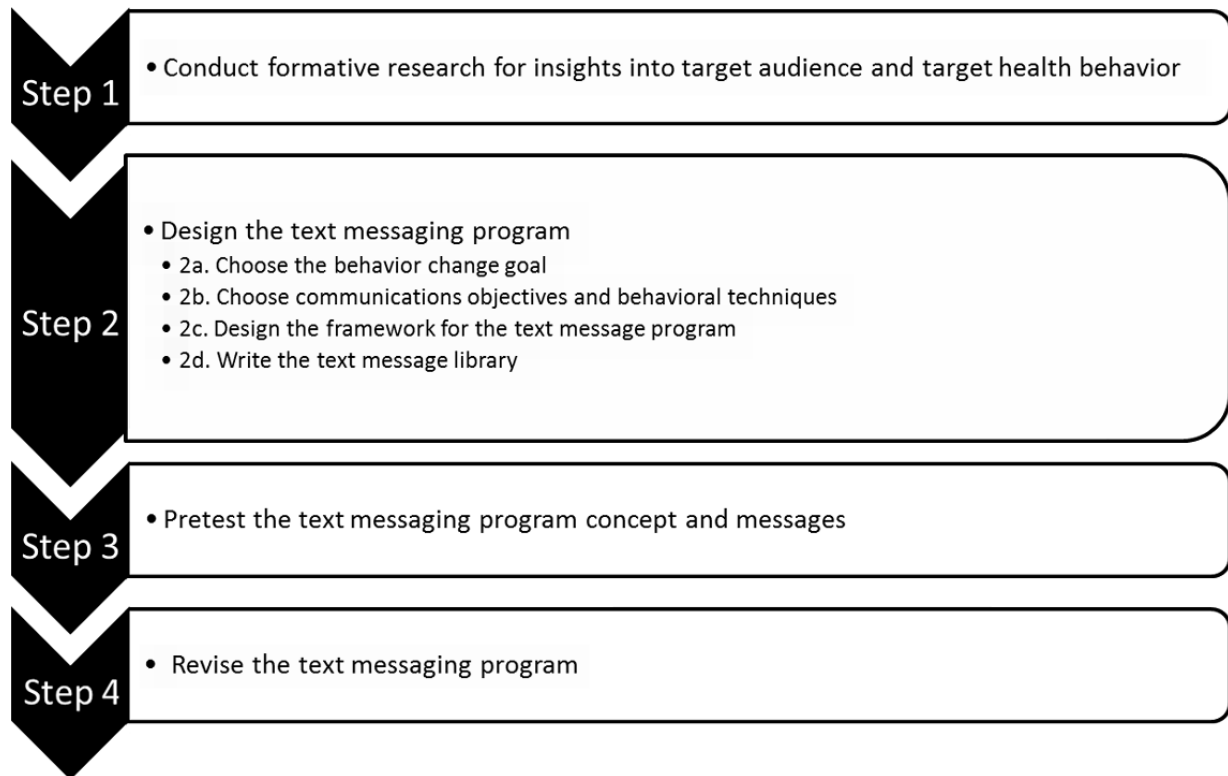
Within the sphere of mHealth interventions, there is growing evidence supporting the success of text-message-based programs (also known as SMS) in modifying health behaviors [6-9]. With >97% of Americans currently owning some type of cell phone and the prevalence of smartphone ownership having increased from 35% to 85% in the last 10 years [10], text messaging has become a staple mode of communication for most people in the modern world. Using a platform already embedded in most individuals' routines, text messages eliminate the need for additional equipment or substantial behavior change. This universal reach and familiarity not only enhances patient engagement but also bridges the gap for underserved communities, thus playing a pivotal role in reducing health disparities [11-13]. Their omnipresent nature, readership and engagement advantage, and the ability to mirror the conversational tone of in-person counseling all underscore the unique value of text messaging in contemporary health interventions [14,15].

Content Creation for Health Care

Content has been described as “the central driver of behavior change” in interventions [16], and its thoughtful incorporation through modalities like text [17,18], imagery [19,20], and other media [21,22], is key to effective intervention design. For brief message interventions in particular, textual content serves not only as a vessel for information, but also as the critical and emotional linchpin motivating behavior change. When we consider interventions designed to induce change, clarity in the content creation process becomes indispensable. It provides a coherent road map for both practitioners and researchers, ensuring that the outcomes of the intervention—successful or not—can be understood, dissected, and refined. Furthermore, a transparent process of content creation not only bolsters the effectiveness of an intervention but also builds trust within the broader scientific community, allowing for constructive critiques, replication of studies, and meaningful advancements in the field.

The conventional SMS intervention development pipeline, shown in Figure 1, consists of the following parts: 1) formative research into the problem setting, behavior, and target population; 2) the establishment of the chosen theoretical framework and development of content; 3) a necessary review of the created content for quality assurance, safety, and research standards, as well as a pretest to gauge initial user feedback on the messages; and finally, a revision of messages based on accumulated feedback [23]. In this study, we address the second step of intervention design—the creation of content within a scientific and theoretical framework—due to its complexity and implication for specialists outside the traditional research team. Content designers can be used to greatly enhance the quality and efficacy of content for health care interventions; however, their involvement is often limited due to monetary, time, and resource constraints on the research team. Consequently, researchers are frequently tasked with taking on the roles of content designers themselves.

Figure 1. Text messaging program development pipeline (reproduced from from Abrams et al [23], which is published under Creative Commons Attribution 4.0 International License [24]).



However, despite the involvement of the research team in content design and the critical importance of content, there exists a conspicuous opacity surrounding the creation of content in health care interventions. Numerous published works fall short in delineating the intricacies of their content creation processes while simultaneously withholding disclosure of their final message data sets, leaving a void in our understanding of how preliminary findings or formative research become translated into the finalized intervention—an omission that has led to the development process being described as a black box [25]. This lack of transparency is especially concerning given the tendency of some researchers to view text messaging as the intervention itself rather than just the means of delivery [26]. In those instances where the message creation process is disclosed, it often reveals a narrow involvement, typically limited to a few individuals within the intervention team [23,27]. Their varied levels of expertise in content design and differing perceptions of what constitutes “good content” can, as a result, lead to vast inconsistencies in outcomes that could be mistakenly attributed to other metrics like participant demographics, study duration, message volume, or the theoretical techniques used instead of the more crucial variable: the nature and quality of the content itself. In addition, this limited participation in content drafting tends to perpetuate familiar methodologies, sidelining innovative approaches that could potentially address persistent challenges like medication adherence [28]. Such exclusions not only hinder academic progress but could also inadvertently reduce the efficacy of interventions. When the foundation—content creation—is not soundly built with a clear and shared understanding of its underpinnings, it runs the risk of diluting the potential positive outcomes of the intervention. As brief message interventions continue to increase in number,

complexity, and scope, the need for innovative and transparent approaches to content creation grows even greater.

Generative AI With LLMs

Faced with the complexities and opacity of content creation, generative AI offers a promising solution to unveiling this enigmatic “black box.” Recent advancements in the development of LLMs using the transformer architecture [29] have brought about a revolutionary change in natural language processing. Unlike earlier models that process text sequentially, transformer models use a technique known as “self-attention” to analyze and draw connections between different parts of input data simultaneously. By converting text into corresponding numerical representations called embeddings, transformers can process language data with exceptional accuracy and speed. Furthermore, being pretrained on vast corpora of web-text data, these LLMs are not only equipped to simulate human conversations but also excel as versatile tools across a spectrum of nuanced tasks, such as question answering, writing support, translation, coding, and more [30-37].

Though the concept of data generation using LLMs is not novel in itself [38-40], the accessibility and enhanced generative capabilities of contemporary large-scale pretrained models like those in the GPT series have magnified their impact and broadened their potential applications [41,42]. With up to hundreds of billions of parameters [32,43], these models excel at rapidly generating vast quantities of contextually appropriate content, streamlining the traditionally painstaking process of manual drafting while simultaneously enhancing adaptability across diverse domains and sectors.

An integral aspect of effective LLM use lies in the art and science of prompt engineering. A prompt is any input given to

an LLM that influences the nature of the LLM's output [44]. Prompts are often given as sets of instructions or requests that establish the rules and guidelines of the conversation. Through prompt engineering, the context of the conversation can be strategically structured to direct the LLM to process relevant information and shape the desired form and content of its output [45]. This process is pivotal in refining and enhancing the capabilities of generative models and allows for the generation of more precise and relevant responses, which is especially imperative in complex fields like health care where the accuracy of information is essential.

Prompt engineering for LLMs is appropriate for the preliminary design of health care interventions for many reasons. First, LLMs can rapidly generate vast amounts of content, effectively reducing both the time and costs typically required for intervention development. This efficiency may allow researchers to allocate resources more appropriately, diverting their energies toward other critical aspects of the project while enabling the exploration of diverse content approaches that were previously considered daunting or impractical. Moreover, LLMs serve as a vital aid to researchers who may not have an extensive background in content design. By providing large amounts of well-written, contextually tailored content, these models offer a structured foundation that researchers can then build upon and further customize during the content review process while avoiding the overwhelm of "blank page paralysis" commonly inherent to creative tasks [46].

Perhaps most significantly for health care intervention research, the application of generative AI introduces a revolutionary level of transparency into the content creation process. By leveraging generative AI models as configurable tools, researchers gain access to a more standardized and reproducible approach for content design. This is primarily enabled through the adjustment of key parameters, such as the "temperature" setting, which are essential for tailoring the models' outputs to specific needs [47,48]. A lower temperature results in more predictable and conservative outputs, while a higher temperature allows for increased variability and creativity in responses. Such configurability not only ensures reproducibility and accessibility but also allows for the establishment of standardized writing styles for health care interventions by minimizing the influence of tone, style, and other confounding variables. With a clearer understanding of the content generation process, researchers are better able to create content at scale, refine content with confidence, and make informed decisions that ultimately enhance the overall efficacy and impact of health care interventions.

Medication Adherence for Type 2 Diabetes

We have chosen the setting of medication adherence for people with type 2 diabetes for our case study on the use of generative AI in health care interventions. Diabetes mellitus currently affects more than 415 million individuals worldwide, with an overwhelming 90% of these instances being attributed to type 2 diabetes [49,50]. Type 2 diabetes is often managed through a combination of dietary modifications, increased physical activity, and the consistent use of oral glucose-lowering medications. However, while oral antidiabetic medications are

often critical to the management of type 2 diabetes, poor adherence to these medications is alarmingly common, with studies suggesting an average adherence rate of only 58% [51,52]. Recent attempts to address this issue have produced mixed results. Notably, a comprehensive review [53] of 182 randomized controlled trials focusing on interventions to improve medication adherence revealed that the evidence supporting their efficacy is largely unconvincing, despite many randomized controlled trials included in the review being extremely time- and resource-intensive. Consequently, such methods are challenging to scale and integrate into routine clinical settings. The paradoxical observation is that the increased complexity and costs of in-person, counseling-style intervention design might not directly lead to better adherence rates, resulting in a pressing need for more innovative, cost-effective, and scalable strategies.

In light of these concerns, SMS-based interventions have emerged as a promising avenue. These brief messaging interventions have previously demonstrated efficacy in promoting various health care behaviors [9,54-57]. Specifically in the domain of type 2 diabetes, interventions based exclusively on messaging [58-60] have shown encouraging results in enhancing medication adherence, though these findings are drawn from a limited number of trials and are not uniformly conclusive [61]. Furthermore, a limitation echoed in these studies is the notable absence of explicit theoretical frameworks guiding the interventions. For these SMS-based interventions to realize their full potential, it is paramount that they are founded on solid theoretical and technical bases, as adopting such an approach ensures that the behavioral mechanisms driving adherence are addressed effectively.

Behavior Change Techniques

Described as the "active ingredients" of an intervention, behavior change techniques (BCTs) epitomize the most fundamental, replicable, and observable elements designed to modify the processes that regulate behavior [62,63]. To translate these strategies into a unified language, a taxonomy encompassing 93 BCTs organized into 16 groups was developed to guide behavior change interventions [64]. This standardization not only aids in replicating and optimizing strategies across various health behaviors but also enhances the comparability of research outcomes. By establishing which techniques are most effective under specific conditions, the taxonomy serves as a valuable resource for researchers and practitioners to select evidence-based approaches tailored to improving behavioral outcomes specific to their patient populations.

However, despite the taxonomy's pivotal role in unifying terminology and subsequently facilitating more comprehensive correlations across behavior change interventions, the application of these BCTs in the realm of message-focused diabetes self-management research remains limited. Among the 93 BCTs outlined in the version 1 taxonomy, only a fraction has been used in published reports for this particular setting [65,66], despite evidence suggesting that interventions using more BCTs typically exert more substantial behavioral effects than those with fewer BCTs [67].

In light of this discrepancy, a comprehensive systematic review of systematic reviews was undertaken to quantitatively pinpoint various BCTs associated with medication adherence across chronic physical health conditions and qualitatively assess them in the context of type 2 diabetes [68]. Overall, the systematic review identified 46 BCTs pertinent to medication adherence in type 2 diabetes that can be used to develop direct messages for mobile devices to improve adherence among patients while simultaneously breaking down the various theoretical constructs (ie, variables from theories targeted by interventions) and mechanisms underlying specific behavioral strategies (ie, techniques not exclusively anchored to one theory but incorporated in interventions due to their predictive value in behavior). Therefore, from this systematic review, there emerges a robust theoretical foundation ripe for practical applications and explicitly suitable for crafting a bank of messages tailored for medication adherence among patients with type 2 diabetes.

Methods

Overview

This paper describes the use of generative AI to develop messages for patients with type 2 diabetes. When using generative AI for nuanced content creation tasks, understanding the context, requirements, and restrictions of the desired content becomes pivotal before initiating the development of a prompt.

Context, Requirements, and Restrictions

Background

This section outlines the key theoretical and technical considerations of our study. Theoretically, we base our content on a preexisting systematic review and widely recognized content design standards to ensure appropriate selection of BCTs and address health disparities through standardized tone and readability. Technically, our focus is on the necessary constraints of SMS delivery systems and the use of a BCT database, which combines findings from the systematic review with fields from the BCT taxonomy to be used conjointly for prompt construction. The following subsections provide detailed insights into each of these aspects.

Problem Setting

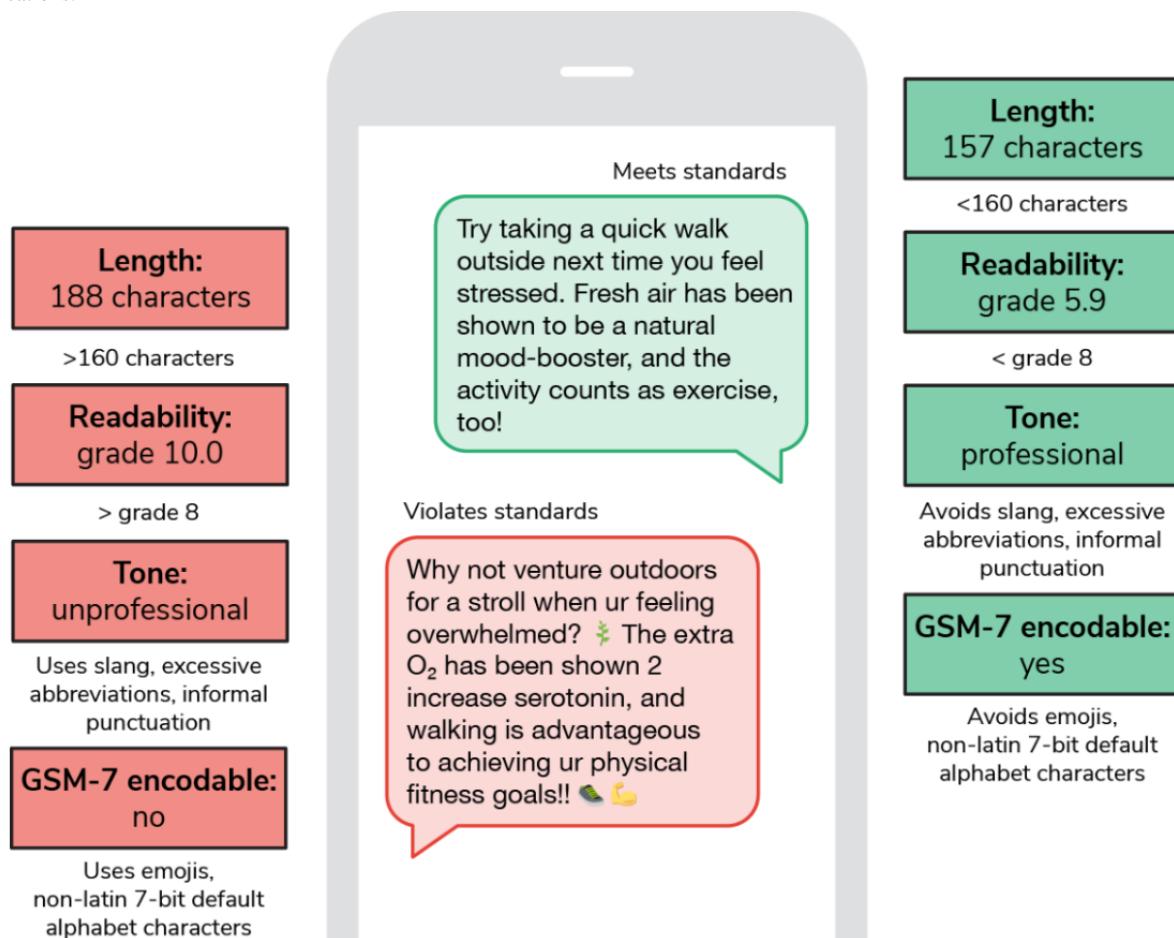
Our setting is based on a rapid systematic review [68] identifying the theoretical constructs and behavioral strategies associated with medication adherence in people with type 2 diabetes and mapping them onto the BCT version 1 taxonomy [64]. The review was done in 2 stages: first, the quantitative review examined interventions and predictors of medication adherence, and second, the qualitative review focused on patients' perceptions, beliefs, and decision-making related to medication adherence specifically for type 2 diabetes. Through this review, 20 theoretical constructs, 19 behavioral strategies, and 46 BCTs were identified as suitable for the content of brief messages to be delivered through mobile devices, which serves as a strong theoretical and scientific underpinning for determining the BCTs and communication objectives used in the content.

Note that the selection of elements used as a theoretical framework in this case study serves as a mere illustration of how one could transform the theoretical framework provided by the research team into a generative AI context. In other applications, the specific information at hand will differ, but the process of integrating such information into prompts may adhere to a comparable methodology.

SMS Standards and Limitations

Messages should ideally be 160 characters (including spaces) or less to be delivered as a single text message to a mobile phone and must consist of only Global System for Mobile Communications (GSM-7)-encodable characters (Figure 2). While some modern smartphones and mobile phone networks allow for message concatenation, enabling longer messages to be sent, requiring smartphone ownership for engagement in health care interventions has been shown to increase health disparities [69]. Thus, the restriction to the 160 characters encodable in GSM-7 has been used in this paper because it is the most standard restriction for SMS-based programs and allows for the greatest number of successful and predictable deliveries to participants.

Figure 2. An example of messages violating (left) and meeting (right) SMS and content design standards. GSM-7: Global System for Mobile Communications.



Content Design Standards

As one must understand a message to be moved by it, literacy demands are a key focus in content design. Messages constructed using shorter words and sentences can cater to a wider range of literacy levels than those using advanced vocabularies and complicated sentence structures. While there are several metrics one might use to evaluate the complexity of a given text [70-73], to ensure accessibility and readability of the generated messages, they were assessed postcreation using the Flesch-Kincaid Grade Level Test due to its widespread use in practice and ease of implementation. The goal reading level is set within or below an 8th-grade level, which is considered the maximum recommended reading level for general adult audiences [74].

In addition, while text messages often carry an informal and conversational tone, in a health care context, even teenage audiences expect there to be a nuanced balance between the relaxed nature of the medium and the professional voice expected from a credible source [75]. Consequently, our messages are designed to avoid the use of slang, excessive abbreviations, or overly informal punctuation. At the same time, messages should convey warmth and friendliness, mirroring the knowledgeable tone of a health care professional with the approachability of a well-informed friend. Figure 2 shows a demonstration of appropriate and inappropriate content design.

BCT Database

To ensure consistency and replicability in the development of messages, we create a standardized database of the 46 BCTs selected based on the needs of the given setting as identified in the systematic review of brief message content [68]. The database contains comprehensive information on each BCT drawn from both the systematic review and the BCT Taxonomy version 1 [64]. By centralizing this data in one location, we can create uniform user prompts that are easily adjustable. This flexibility allows for structural modifications, the inclusion or omission of different fields, and swift adaptation if further curation of BCTs is required, thus ensuring both consistency and flexibility in the development of targeted health care interventions. There are six database fields:

1. Number—the number assigned to the BCT [64]
2. Label—the name of the BCT [64]
3. Definition—the definition of the BCT [64]
4. Examples—available examples of the BCT [64]
5. Theoretical constructs—theoretical constructs mapped to the BCT [68]
6. Behavioral strategies—behavioral strategies mapped to the BCT [68]

The final table containing the BCT database can be found in [Multimedia Appendix 1](#).

Technical Setup

To generate textual content, a pretrained LLM is needed. While a variety of options currently exist, both proprietary (eg, GPT [32] and LaMDA [76]) and open-source (eg, Orca [77] and Llama 2 [78]), we use GPT for this particular project. As one of the most advanced and widely recognized models in the field of AI-driven language generation [41,42,79,80], GPT benefits from an extensive body of research and a thriving community of developers.

In this work, we use the gpt-3.5-turbo-0301 model through OpenAI application programming interface (API) [47] calls to generate health care messages and use the chat completion API to communicate with the model. While the use of the completion functionality might appear to be more suitable for a single-prompt interaction, we observed superior results through the chat function during initial testing, and therefore continued development in a chat setting. Moreover, this choice aligns with the practical recommendations provided by OpenAI [48].

As our experiments are of an illustrative nature, we mostly use default parameter values for the API calls. In more nuanced use cases, these values could be tweaked by the prompt engineer to further tailor the output of the model to comply with the application, but adjustments were unnecessary for our use case. However, to ensure reproducibility of the presented results, we globally set temperature equal to 0, even though in practice one may obtain better outcomes by setting a positive temperature and rerunning the same query until a more satisfactory result is achieved. For instance, in ChatGPT, the value of the temperature is set to 0.7, which allows for more varied, human-like responses.

Message generation and analysis are performed in a Jupyter notebook using Python 3.8 on a consumer-grade laptop. The source code is included in [Multimedia Appendix 1](#).

Prompt Engineering

For this work, we consider single-prompt chat completion where the messages parameter contains 2 roles—"system" and "user"—and their corresponding "content." The conversation begins with an initial system prompt, followed by a prompt from the user. The interaction concludes with a response from GPT, which provides the generated output. To enhance the performance of the model for a task with many constraints (in our case, these included length, complexity, style, and BCT incorporation), attributed prompt design [81] has been used for both the system and user roles.

The system role provides general context and behavior instructions to the assistant. It is used to explain the setting, rules, parameters, and personas of each participant in the conversation.

The content of our system prompt is given in [Textbox 1](#) and consists of four main components:

1. **Setting**— this establishes the general setting of the conversation, that is, the designated roles of "user" (as behavioral scientist) and "assistant" (as diabetes specialist), and the goal of the interaction (to construct messages encouraging meditation adherence).
2. **Style rules**—these are guidelines on style to be used by the assistant when constructing messages. In this case, style rules focus mostly on the personality of the messages, in addition to limitations on length, complexity, and uniqueness.
3. **BCT rules**—these are guidelines on the incorporation of BCTs to be used by the assistant when constructing messages. BCT rules explain the importance of the BCT and give directions for use.
4. **Task**—this combines the previous 3 sections into a single, condensed statement defining the particular task being asked of the "assistant" role.

The user role begins the conversation by providing the first interaction to which the assistant role can respond. In our setting, the user role has been defined through the system prompt as "behavioral scientist," and reflects a templated version of the BCT database to provide the assistant with the selected BCT and its corresponding information.

The structure of our user prompt is given in [Textbox 2](#), where the tokenized attributes are replaced with their corresponding values from the BCT database for each query. The five attributes used in our prompts are as follows:

1. `bct_label`: the name of the selected BCT [64], prepended by the label "BCT: "
2. `bct_definition`: the definition of the selected BCT [64], formatted in line with the `bct_label` following an equal sign (=)
3. `bct_examples`: if available, examples of the selected BCT [64], formatted as a new line prepended by the phrase "For example, "
4. `bct_theoretical_constructs`: if available, the theoretical constructs corresponding to the selected BCT [68], separated by 2 line breaks and prepended by the label "Theoretical Constructs: "
5. `bct_behavioral_strategies`: if available, the behavioral strategies corresponding to the selected BCT [68], formatted as a new line and prepended by the label "Behavioral Strategies: "

Results are delivered through the "assistant" role, which is the content generated by the chosen GPT model in response to each particular combination of system and user inputs. To maintain the consistency of the presented results, we postprocess the model output by stripping quotation marks and standardizing the message separation to a single line break.

Textbox 1. Attributed system prompt used for message generation.

You are a Diabetes Specialist encouraging medication adherence in people with type 2 diabetes via brief messages.

Your messages are informed by different Behavior Change Techniques (BCTs).

I am a Behavioral Scientist who will describe the BCT you should use to frame your messages encouraging medication adherence.

Messages should be friendly and positive, but also professional, super short, and to-the-point. You are limited on space. Messages should be written at the reading level of an eighth grader. Word choice should be short and simple so everyone can understand. Every message must be entirely unique from all others in both language and structure.

The BCT I will provide is {bct_label}. It is the most important thing, and it is very nuanced. Messages must intelligently use {bct_label} to encourage medication adherence. All messages must prioritize {bct_label} over everything else. DO NOT write any part of the user message verbatim -- the BCT, theoretical constructs, and behavioral strategies are a secret.

Task: You will use these sets of rules to construct 25 diverse messages that use {bct_label} to increase medication adherence for recipients with type 2 diabetes.

Textbox 2. Attributed user prompt used for message generation.

BCT: {bct_label} = {bct_definition}

For example, {bct_examples}

Theoretical Constructs: {bct_theoretical_constructs}

Behavioral Strategies: {bct_behavioral_strategies}

{bct_label} is critical to each creative, chatty message.

Ethical Considerations

This research focused on the development of content for SMS interventions using generative AI, which did not require the collection or analysis of personal data or direct engagement with individuals. According to federal guidelines from the US Department of Health and Human Services, known as the “Common Rule” (45 CFR 46) [82], ethics board approval is required only for research on human subjects that entails obtaining data through interaction with individuals or the use of identifiable private information. Since this study did not meet these criteria, an ethics review was not applicable. The generated content and associated processes were evaluated for adherence to ethical standards in the context of AI-generated material and were designed to be transparent, reproducible, and free of harm.

Results

Overview

Using the attributed system and user prompts, 25 messages are generated for each of the 46 selected BCTs, resulting in 1150 BCT-informed messages. A total of 45,905 tokens costing US \$0.07 were used for the entire data set generation, with an average of 998 tokens used and US \$0.0015 expenditure per BCT. Approximately US \$10 was spent for the entire prompt development process, which required over 300 iterations of the prompt. The first 5 generated messages for the first 6 BCTs are presented in [Table 1](#), with the full data set made available in [Multimedia Appendix 1](#).

We analyze the generated messages to evaluate their overall adherence to the required metrics. Specifically, we consider the following parameters: message length, message readability, and diversity of messages and their BCT alignments compared with an existing message data set.

Table 1. Examples of messages generated from the user query.

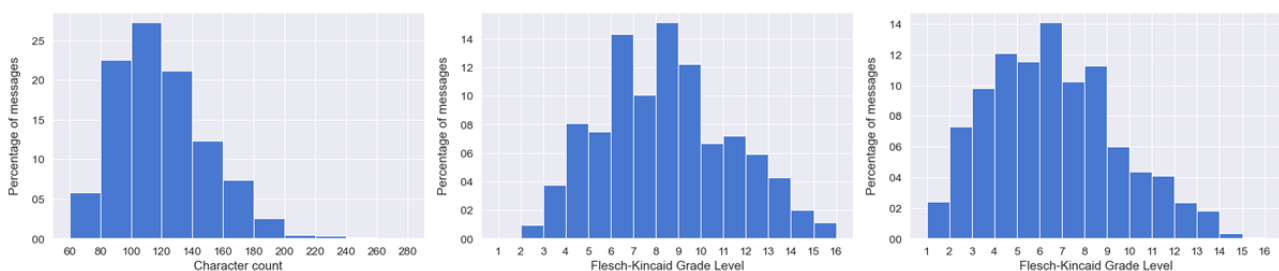
User query	Message generated
1.2 Problem solving	<ul style="list-style-type: none"> Feeling overwhelmed by your medication regimen? Let’s work together to simplify it and make it more manageable. Are you experiencing any side effects from your medication? Let’s problem solve and find ways to cope with them. Let’s identify any barriers preventing you from taking your medication as prescribed and develop strategies to overcome them. Having trouble remembering to take your medication? Let’s problem solve and find ways to make it a habit. Are you concerned about the cost of your medication? Let’s problem solve and explore options for financial assistance.
1.4 Action planning	<ul style="list-style-type: none"> Set a reminder on your phone to take your medication every day at the same time. This will help you remember to take it consistently. Think about the best time of day for you to take your medication and plan to take it at that time every day. Create a checklist of all the medications you need to take and when to take them. This will help you stay organized and on track. Plan to take your medication with a meal or snack to help prevent any stomach upset. Make a plan to refill your medication before you run out. This will ensure you always have it on hand.
2.3 Self-monitoring of behavior	<ul style="list-style-type: none"> Keep track of your medication intake every day to ensure you are staying on track with your treatment plan. Make a habit of checking your blood sugar levels regularly to monitor the effectiveness of your medication. Keep a log of any symptoms you experience to help your health care provider adjust your medication as needed. Set reminders on your phone to take your medication at the same time every day to establish a routine. Use a medication tracker app to monitor your medication intake and set reminders for refills.
2.4 Self-monitoring of outcome(s) of behavior	<ul style="list-style-type: none"> Keep track of your blood sugar levels every day to see how your medication is working for you. By monitoring your blood sugar levels, you can see the positive impact your medication is having on your health. Make a habit of recording your blood sugar levels to stay on top of your medication regimen. Keep a log of your blood sugar levels to help you stay motivated to take your medication as prescribed. By tracking your blood sugar levels, you can see how your medication is helping you manage your diabetes.
2.7 Feedback on outcome(s) of behavior	<ul style="list-style-type: none"> Great job! By taking your medication as prescribed, you are helping to manage your diabetes and prevent complications. Keep up the good work! Taking your medication regularly can help lower your blood sugar levels and reduce your risk of long-term health problems. You are doing an amazing job! Remember, taking your medication as directed can help you feel better and stay healthy. Congratulations! By sticking to your medication regimen, you are taking control of your diabetes and improving your overall health. Way to go! Taking your medication on time and as prescribed can help you avoid complications and live a longer, healthier life.

Message Length

Messages should be 160 characters or less. In total, 1034 of the 1150 (89.91%) total generated messages satisfy this requirement

(Figure 3). The average length of all generated messages is 119 characters.

Figure 3. Length (left) and readability (before adjustment, center; after adjustment, right) of the generated messages.



Message Readability

Messages should fall within or below an 8th-grade level on the Flesch-Kincaid Grade Level Test. In total, 688 of the 1150 (59.83%) total generated messages satisfy this requirement with no alterations for our setting (Figure 3). The average grade level of all generated messages is 8.4.

While this is the initial score for all messages, it is critical to note that the nonnegotiable and unsubstitutable word

“medication” is considered complex due to its character length (10 characters) and number of syllables (4 syllables). However, for our setting, it is assumed that a population prescribed diabetes-management medications will be cognizant of the word “medication,” and it will thus not pose the same complexity barrier in our context as it might in other applications. Therefore, when the word “medication” is ignored during the readability calculation, 928 out of the 1150 (80.7%) total messages satisfy the readability requirement, with an average grade of 6.5 (Figure

3)—a closer, more accurate metric of complexity for our particular use case.

Message Diversity

Overview

To evaluate the diversity of the generated messages, we compare them to the largest publicly available data set of SMS health care communications using BCTs to address behaviors surrounding diabetes [27]. We use pretrained natural language processing systems to compute the embeddings for each set of messages and compare their distribution.

It is important to note that due to the general opacity surrounding message creation for brief message interventions and the resulting lack of publicly available data sets, the study [27] we use for comparison is similar in theoretical framework used (BCTs) and general condition (diabetes), but different in population (individuals with prediabetes vs diagnosed diabetics), health behaviors addressed (diet and physical activity vs medication adherence), and size of the data set (124 vs 1150).

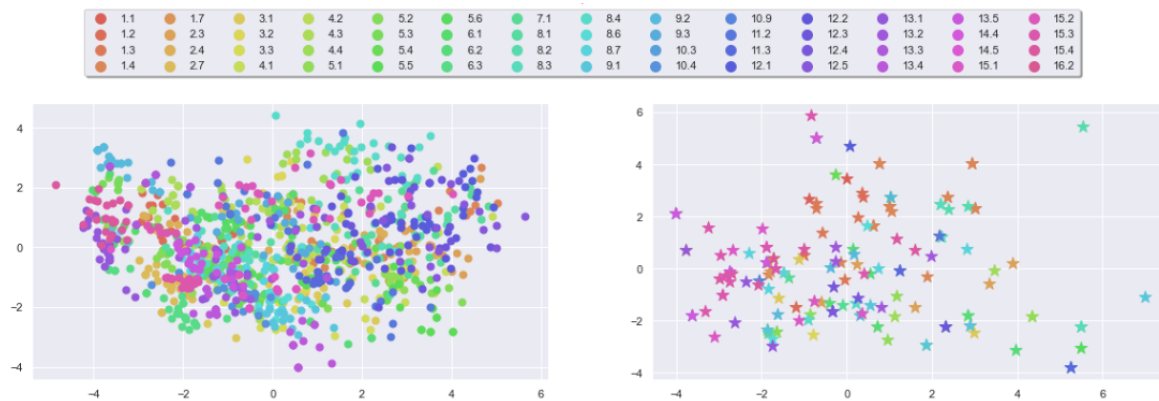
Also note that some of the messages in the comparison data set [27] are coded for multiple BCTs. In such cases, we duplicate the message and assign each variation a single BCT to be consistent with our single-BCT-per-message mapping, resulting in a comparison data set consisting of 169 total messages.

BERT Embeddings and Principal Component Analysis Projection

We use BERT [83] to compute message embeddings through the bert-base-uncased model available through the Hugging Face Inference API [84]. For any message $x \in X$ its BERT embedding vector $\text{emb}(x)$ is given as $\text{emb}(x) \in \mathbb{R}^{768}$.

For each message, we compute its 768-dimensional embedding vector and then project it onto a 2D plane using principal component analysis (PCA) [85] (Figure 4). We note that the distributions of embeddings in both data sets are comparable, with embeddings being spread throughout the latent space without clustering per BCT, which indicates the presence of nontrivial semantic diversity.

Figure 4. Principal component analysis projection of BERT embeddings of messages: ours (left) and comparison (right).



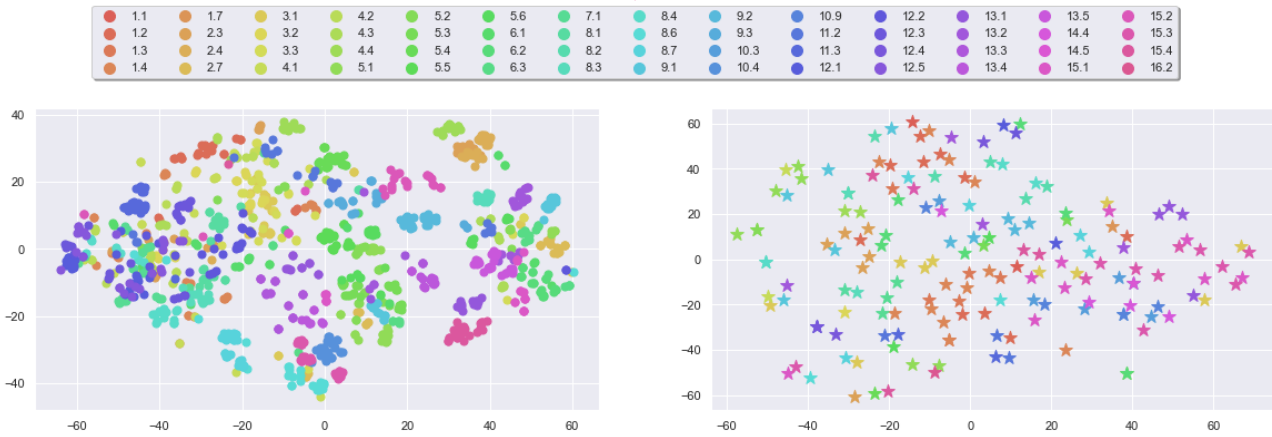
ADA Embeddings and t-Distributed Stochastic Neighbor Embedding Visualization

We use ADA [86] to compute message embeddings through the text-embedding-ada-002 model available through the OpenAI Embeddings API [87]. For any message $x \in X$, its ADA embedding vector $\text{emb}(x)$ is given as

$$\text{emb}(x) \in S^{1535} \subset \mathbb{R}^{1536},$$

where S^{1535} denotes the unit sphere in \mathbb{R}^{1536} . Because the embeddings computed by ADA are given as points on the unit sphere of the latent space, it does not seem sensible to use a linear projector like PCA; instead, we use t-distributed stochastic neighbor embedding (t-SNE) [88] to perform nonlinear dimensionality reduction. For each message, we compute its 1536-dimensional embedding vector and then embed them into a 2D plane using t-SNE (Figure 5). We note that the distributions of embeddings of both data sets are comparable, with messages corresponding to the same BCT being positioned closely.

Figure 5. t-Distributed stochastic neighbor embedding visualization of ADA embeddings of messages: ours (left) and comparison (right).



Cross-Comparison of Data Sets

In this section, we aim to provide a more head-to-head comparison between the 2 data sets. To achieve such a comparison, 2 major differences in the data sets must be addressed: *distribution* (ie, the BCTs and number of messages per BCT) and *objective* (ie, the setting for which the messages are written).

To match the BCT-message *distribution* in 2 data sets, we first select the messages corresponding to the BCTs present in both data sets (Multimedia Appendix 1). Then, we check the number of messages available for each BCT in the comparison data set (169 messages mapped onto 41 BCTs) and take the same number of messages for each corresponding BCT from ours (1150 messages mapped onto 46 BCTs). This results in 2 sets of 135 messages spread across 31 BCTs (Multimedia Appendix 1).

While matching the objective is barely feasible, as it involves changing the semantic structure of each message, we attempt to nullify this difference by averaging the embeddings over each

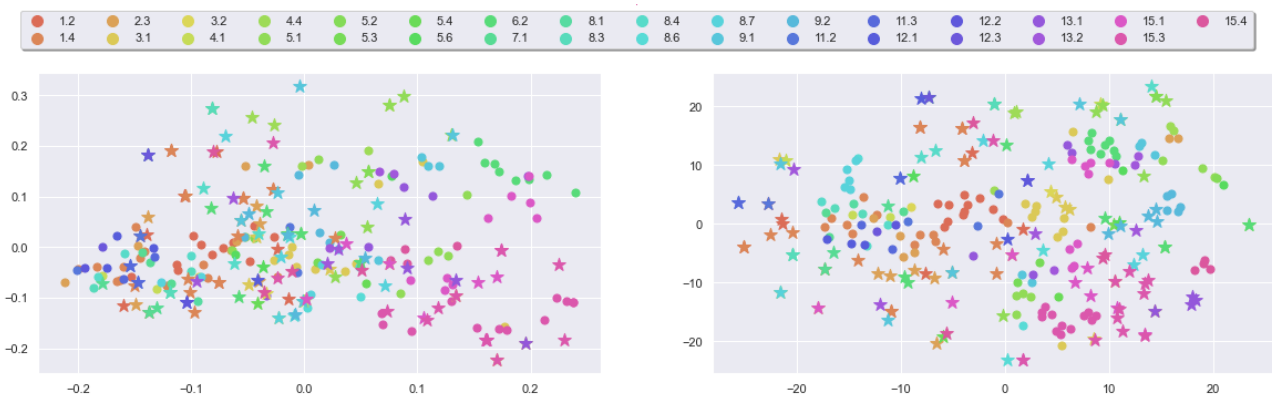
data set. Concretely, for each message x from the data set X we compute its representation $r(x)$ by taking the ADA embedding $emb(x)$ and centering it as,

$$r(x) = emb(x) - \frac{1}{|X|} \sum_{x \in X} emb(x) \tag{1}$$

where X denotes the set of all messages from this data set and $|X|$ denotes its cardinality. This modification is proposed in [89] and is prefaced on the assumption that the embedding $emb(x)$ contains sufficient semantic information about the message $x \in X$, and thus the average of the embedding vectors over the data set represents the information that unifies all the messages, that is, the *objective*. By subtracting the average, the representation $r(x)$ still contains the information that is specific to this particular message, that is, the semantic structure and the BCT.

Computing representations (equation 1) for each message in both data sets allows us to directly compare the 2 data sets through the PCA and t-SNE projection of the message representations, shown in Figure 6.

Figure 6. Projections of normalized ADA embeddings of messages: principal component analysis (PCA; left) and t-distributed stochastic neighbor embedding (t-SNE; right).



Moreover, we compute the relevance between BCT encodings in the 2 data sets by averaging message representations corresponding to each BCT and then taking an inner product, that is,

$$r(x_1) \cdot r(x_2)$$

where $\cdot, \cdot : \mathbb{R}^{1536} \times \mathbb{R}^{1536} \rightarrow \mathbb{R}$ denotes the inner product, X_1/X_2 is the set of messages corresponding to BCT bct_1/bct_2 , respectively, and $r(x)$ is the numeric representation of the message x computed via equation 1. The resulting 31×31

heatmap can be found in [Multimedia Appendix 1](#). The obtained relevancies can be used to evaluate the alignment of BCTs between 2 data sets, which result in a top-5 accuracy of 67% and a top-10 accuracy of 87%. Even though such an approach is a bit heuristic, we observe that the representations of the messages from both data sets are distributed similarly, often with messages corresponding to the same BCT being close to one another. This observation provides grounds to contend that the diversity of the messages generated by our approach is comparable to those previously created by researchers for practical, real-world applications.

Discussion

Overview

In this paper, we propose a novel approach to creating behaviorally informed content for brief message interventions. Using the setting of medication adherence for people with type 2 diabetes, we use a pretrained LLM to develop a bank of text messages based on BCTs curated in a recent systematic review [68]. This work is intended to act as a blueprint for future research to create a more transparent, replicable, and scientifically rigorous look into the content creation process for brief message interventions and serve as a starting point for subsequent studies to analyze the safety, efficacy, and viability of AI-generated messages.

Principal Findings

In this paper, we show the potential of generative AI as a tool for transparent and replicable content creation in brief message interventions. Inspired by a list of 46 BCTs and their corresponding theoretical constructs and behavioral strategies, we engineer attributed system and user prompts for GPT to generate 25 messages for each of the 46 BCTs, for a total of 1150 messages in the specific setting of medication adherence for type 2 diabetes. Our findings reveal that a significant majority of the generated messages were compliant with both message length and complexity considerations (1034/1150, 89.91% and 928/1150, 80.7%, respectively), making them well-suited for SMS-style interventions.

The diversity of generated messages is analyzed through the distributions of their embedding vectors with 2 popular pretrained natural language processing systems: BERT and ADA. The generated messages showcase a diversity in message content that is comparable with an existing publicly available data set of brief messages and reflects similar distributions among BCTs from the comparison data set while also maintaining variability between messages of the same BCT, thereby demonstrating the capability for generative AI to craft a plethora of unique and contextually relevant communications with only a very standardized change in input between BCTs. The data set and source code for message generation and analysis are available in [Multimedia Appendix 1](#).

Technical Limitations

As generative AI is an extremely new field growing at a rapid rate, new algorithms and “versions” of LLMs are being released regularly. These updates can often fundamentally change the assistant output generated by the same system and user inputs,

leading to a lack of consistency in experiments conducted with the same prompt over a long period of time. Especially for users of chatbot-style LLMs like ChatGPT, these updates can come suddenly and without permission, making research on such platforms difficult. While this issue is somewhat mitigated using an API (which generally does not force immediate adoption of the newest models), many current LLMs will eventually be depreciated, albeit at a more gradual rate. The only true mitigation of this limitation is the use of open-source models (such as, for example, Orca [77] and Llama 2 [78]) that can be fully downloaded and deployed on the client side; however, this comes at the cost of the technical proficiency required to set up such a system. Therefore, while prompts like ours serve as great examples of attributes to consider and the language one might use when constructing a prompt for a particular setting, the definitive construction of a singular, unchanging prompt to support intervention research is generally unfeasible in practice.

Another potential limitation of generative AI for very large-scale message generation is the finite context window size of the given LLM. In our results, we generate 25 messages per BCT, equating to an average of 998 tokens for each interaction, which keeps us well within the current 4000-token context limit bounds of gpt-3.5-turbo-0301. If a larger bank of messages is required, according to our current standardized single-prompt structure, one could feasibly increase the number of generated messages to 100 or more before nearing the limit. However, for extremely large-scale data generation, even a single-prompt interaction will likely be insufficient, and because LLM context is cumulative, this restriction will likely be most prominent for multistep interactions.

It is important to note that our use of a singular templated prompt across multiple BCTs can sometimes fail to capture the nuanced essence of each distinct BCT. This “one-size-fits-all” approach may yield inconsistencies in the quality and accuracy of the generated content for certain BCTs as compared with others. However, prior studies in behavioral science have also suggested that some BCTs may not be capable of being delivered effectively in an SMS format regardless of the method of creation [90], perhaps indicating that difficulties in accurately representing some BCTs may have less to do with the limitations of generative AI and more to do with the inherent complexities of some BCTs in the given setting of brief message interventions.

Generative AI, while versatile, is heavily reliant on the specific language of the prompts provided; thus, when prompts rely heavily on external data fitting into a templated structure, small, seemingly insignificant differences in the style or authorship of that data can potentially affect results [91,92]. For those aiming for AI-generated content that would be used in practice without human review, the design and fine-tuning of dedicated prompts for each unique contextual change (in our case, BCTs and their corresponding information) are likely a necessity. Therefore, while our approach serves as a proof of concept for efficient single-prompt content creation and demonstrates the vast potential of pretrained LLMs in this domain, achieving universally accurate results for all BCTs would necessitate a more granular, tailored approach to prompt

design, attending to the individual nuances and requirements of each BCT within its specific context.

Safety and Ethics in AI

As with any state-of-the-art technology, ethical considerations for the implementation of generative AI are paramount, guided by the core principles of transparency, privacy, accountability, and fairness [93,94]. However, the inherent unpredictability of LLMs becomes acutely significant in health care contexts, where the consequences of misinformation or inappropriate AI-generated suggestions can be dire for patients. Given that ensuring fully reliable, safe, and accurate information generation by LLMs is deemed “fundamentally impossible” [95], human checkpoints become indispensable before, during, and after AI employment. Researchers and domain experts must meticulously review AI-generated content for accuracy, safety, and equity, especially when evaluating its usability for complex patient-facing health care interventions.

It has been demonstrated that diversely attributed, complex prompts can reduce biases [81], but it is imperative to strike a balance, as excessively long prompts may increase the likelihood of undesired model behaviors [96]. In our approach, we acknowledge the crucial role of domain experts in both prompt design and content review, effectively mitigating the risk of malicious actor involvement and enabling the use of longer, more attributed prompts. Generative AI should not be seen as a standalone solution but as a tool that augments and accelerates the work of researchers. While it dramatically enhances efficiency in content creation, the responsibility for upholding rigorous standards of accuracy, safety, and fairness in health care interventions remains firmly with human experts, and it is the symbiotic collaboration between this tool and the human research team that ensures the delivery of ethically sound and clinically effective interventions.

Comparison With Prior Work

To the best of our knowledge, this is the first work to propose the use of generative AI as a tool for content creation in SMS health interventions. However, a similar study [28] detailing a traditional content creation process was undertaken using the same systematic review [68] as a theoretical framework. A workshop was held for content creation and subsequent focus groups and surveys were used for review, resulting in the production of 371 messages informed by the selected BCTs in the context of medication adherence for type 2 diabetes. However, despite efforts toward transparency, this work does not reveal a detailed account of the actual content creation process, and the data set of generated messages has not been made publicly available for review or comparison.

Previous studies have looked at traditional content creation for brief message interventions, with a specific focus on the selection and review of BCTs and their corresponding messages [27,97,98]. More broadly, investigations into mHealth interventions have been a hot topic in health care research for years [11,14,15,26], and several works investigating the specific incorporation of behavioral science into brief message interventions have been previously undertaken with positive results [99-101].

In addition, one-shot, zero-shot, and few-shot approaches to prompt engineering have seen an explosion of interest following the expansion of LLMs within the public and academic mindset, leading to a large body of research on the methods and frameworks of prompt design for a variety of contexts and use cases [45,102,103].

Future Work

This is the first in a series of works detailing the process for responsible and efficient use of generative AI in the development of brief message health care interventions. The next step involves assembling a team of qualified behavioral scientists and other domain experts to conduct in-depth analyses of the generated messages, focusing on their adherence to safety standards, adjustments to meet the technical requirements of an SMS delivery system, a formal review of BCT coding for each message, and general checks that the generated messages meet best practice standards for content design.

Moreover, the development of a subsequent interaction with the model could be used to self-adjust the generated results based on designer feedback. Using a multistep prompting method, the model could, for example, be directed to self-assess for safety and equity considerations, as well as edit more individually for the given use case based on specific critiques provided by the research team. Such iterative developments of the model should necessarily involve rigorous patient testing and feedback—a crucial step in ensuring that the AI-generated content resonates with patients’ experiences and needs to further personalize and refine the developed health care interventions.

Standardizing the realization of individual BCTs within brief message content represents another critical research direction. Many interventions currently withhold both their messages and their content creation processes, potentially introducing unintentional biases and skewed outcomes due to inherent differences in writing styles and other design-related confounding variables. By advocating for the transparency and standardization of content design, we can enhance the research efficacy of interventions by reducing such confounders and further ensuring that the results of such interventions are truly tied to the theoretical frameworks and behaviors being tested.

Finally, the ambitious goal of generating hyperpersonalized health care communications tailored to individual patients becomes a promising possibility with LLMs. Future iterations of content could be further personalized by implementing features like translating content for different languages and localizations; tailoring content for cultural relevance and sensitivity in the use of examples, metaphors, and references; adjusting the complexity levels of the text to cater to different educational backgrounds or cognitive abilities; and providing accessible formats adjusted for individuals with disabilities. While conventional content creation methods struggle with the impossible number of potential messages required for personalized content for every recipient, attributed prompting with generative AI offers the potential to create individualized messages that can significantly enhance patient engagement and outcomes and change the way health care interventions are experienced.

Conclusions

In this study, we explore the practical application of generative AI for content creation in the development of brief message health care interventions. We illustrate the potential of using pretrained LLMs as a tool to aid researchers in the resource-intensive process of content creation by generating a data set of 1150 messages inspired by 46 BCTs selected for the setting of medication adherence for type 2 diabetes. Building on the foundations laid by former health care intervention development studies, this paper differentiates itself by the following:

1. Proposing and demonstrating the use of generative AI with pretrained LLMs for intervention development
2. Detailing the use of state-of-the-art AI tools for prompt engineering and content design processes
3. Providing the largest publicly available data set of messages created for SMS interventions, as well as the first publicly available source code offering fully transparent insight into the content creation process

Ultimately, the value proposition of using generative AI in this domain lies not in the perfection of the initial generated content but in its adaptability and capacity to rapidly produce a multitude of messages that can subsequently be refined and curated by human experts. This combination of AI-driven speed and human-driven supervision presents an efficient, transparent, and scalable method for developing effective and replicable brief message interventions. While follow-up studies are needed to ensure the safety and usability of the generated messages and provide potential refinements to the proposed prompts for individual settings, the use of generative AI in health care intervention development opens new doors for the scalability and potential standardization of content creation within health care intervention design and research. Given the time- and cost-intensive nature of crafting interventions traditionally and the current opacity of the content design process, our study underscores the potential of generative AI as a significant efficiency tool poised to revolutionize the creation of behavior change interventions for medication adherence and beyond.

Acknowledgments

The authors are grateful to Dr Sofya Belova of the Institute of Psychology at the Russian Academy of Sciences and Dr Anton Dereventsov of Klaviyo for their advice and helpful discussions.

Authors' Contributions

RMH's contributions to the study were independent of her former affiliation at Lirio LLC. EL's contributions were independent of her former affiliation at the Institute of Psychology at the Russian Academy of Sciences. During the process of developing this paper, AB transitioned departments at the Skolkovo Institute of Science and Technology from Skoltech Agro to Skoltech AI. The generative artificial intelligence model GPT-3.5 by OpenAI was used to generate the database of messages created and analyzed in this study. The source code and generated messages are made available in [Multimedia Appendix 1](#).

Conflicts of Interest

None declared.

Multimedia Appendix 1

Source code and data tables.

[\[ZIP File \(Zip Archive\), 1513 KB - ai_v3i1e52974_app1.zip\]](#)

References

1. Smith V, Devane D, Begley CM, Clarke M. Methodology in conducting a systematic review of systematic reviews of healthcare interventions. *BMC Med Res Methodol* 2011 Feb 03;11(1):15 [FREE Full text] [doi: [10.1186/1471-2288-11-15](https://doi.org/10.1186/1471-2288-11-15)] [Medline: [21291558](https://pubmed.ncbi.nlm.nih.gov/21291558/)]
2. Ghersi D, Pang T. From Mexico to Mali: four years in the history of clinical trial registration. *J Evid Based Med* 2009 Feb;2(1):1-7. [doi: [10.1111/j.1756-5391.2009.01014.x](https://doi.org/10.1111/j.1756-5391.2009.01014.x)] [Medline: [21348976](https://pubmed.ncbi.nlm.nih.gov/21348976/)]
3. van Heerden A, Tomlinson M, Swartz L. Point of care in your pocket: a research agenda for the field of m-health. *Bull World Health Organ* 2012 May 01;90(5):393-394 [FREE Full text] [doi: [10.2471/BLT.11.099788](https://doi.org/10.2471/BLT.11.099788)] [Medline: [22589575](https://pubmed.ncbi.nlm.nih.gov/22589575/)]
4. Marcolino MS, Oliveira JA, D'Agostino M, Ribeiro AL, Alkmim MB, Novillo-Ortiz D. The impact of mHealth interventions: systematic review of systematic reviews. *JMIR Mhealth Uhealth* 2018 Jan 17;6(1):e23 [FREE Full text] [doi: [10.2196/mhealth.8873](https://doi.org/10.2196/mhealth.8873)] [Medline: [29343463](https://pubmed.ncbi.nlm.nih.gov/29343463/)]
5. Stowell E, Lyson MC, Saksono H, Jimison H, Wurth RC, Pavel M, et al. Designing and evaluating mHealth interventions for vulnerable populations: a systematic review. In: *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 2018 Presented at: CHI '18; April 21-26, 2018; Montreal, QC. [doi: [10.1145/3173574.3173589](https://doi.org/10.1145/3173574.3173589)]
6. Hall AK, Cole-Lewis H, Bernhardt JM. Mobile text messaging for health: a systematic review of reviews. *Annu Rev Public Health* 2015 Mar 18;36:393-415 [FREE Full text] [doi: [10.1146/annurev-publhealth-031914-122855](https://doi.org/10.1146/annurev-publhealth-031914-122855)] [Medline: [25785892](https://pubmed.ncbi.nlm.nih.gov/25785892/)]

7. Head KJ, Noar SM, Iannarino NT, Grant Harrington N. Efficacy of text messaging-based interventions for health promotion: a meta-analysis. *Soc Sci Med* 2013 Nov;97:41-48. [doi: [10.1016/j.socscimed.2013.08.003](https://doi.org/10.1016/j.socscimed.2013.08.003)] [Medline: [24161087](https://pubmed.ncbi.nlm.nih.gov/24161087/)]
8. De Leon E, Fuentes LW, Cohen JE. Characterizing periodic messaging interventions across health behaviors and media: systematic review. *J Med Internet Res* 2014 Mar 25;16(3):e93 [FREE Full text] [doi: [10.2196/jmir.2837](https://doi.org/10.2196/jmir.2837)] [Medline: [24667840](https://pubmed.ncbi.nlm.nih.gov/24667840/)]
9. Armanasco AA, Miller YD, Fjeldsoe BS, Marshall AL. Preventive health behavior change text message interventions: a meta-analysis. *Am J Prev Med* 2017 Mar;52(3):391-402. [doi: [10.1016/j.amepre.2016.10.042](https://doi.org/10.1016/j.amepre.2016.10.042)] [Medline: [28073656](https://pubmed.ncbi.nlm.nih.gov/28073656/)]
10. Mobile fact sheet. Pew Research Center. URL: <https://www.pewresearch.org/internet/fact-sheet/mobile/> [accessed 2023-07-12]
11. Militello LK, Kelly SA, Melnyk BM. Systematic review of text-messaging interventions to promote healthy behaviors in pediatric and adolescent populations: implications for clinical practice and research. *Worldviews Evid Based Nurs* 2012 Apr;9(2):66-77. [doi: [10.1111/j.1741-6787.2011.00239.x](https://doi.org/10.1111/j.1741-6787.2011.00239.x)] [Medline: [22268959](https://pubmed.ncbi.nlm.nih.gov/22268959/)]
12. Koivusilta LK, Lintonen TP, Rimpelä AH. Orientations in adolescent use of information and communication technology: a digital divide by sociodemographic background, educational career, and health. *Scand J Public Health* 2007;35(1):95-103. [doi: [10.1080/14034940600868721](https://doi.org/10.1080/14034940600868721)] [Medline: [17366093](https://pubmed.ncbi.nlm.nih.gov/17366093/)]
13. Faulkner X, Culwin F. When fingers do the talking: a study of text messaging. *Interact Comput* 2005 Mar;17(2):167-185. [doi: [10.1016/j.intcom.2004.11.002](https://doi.org/10.1016/j.intcom.2004.11.002)]
14. Krishna S, Boren SA, Balas EA. Healthcare via cell phones: a systematic review. *Telemed J E Health* 2009 Apr;15(3):231-240. [doi: [10.1089/tmj.2008.0099](https://doi.org/10.1089/tmj.2008.0099)] [Medline: [19382860](https://pubmed.ncbi.nlm.nih.gov/19382860/)]
15. Fjeldsoe BS, Marshall AL, Miller YD. Behavior change interventions delivered by mobile telephone short-message service. *Am J Prev Med* 2009 Feb;36(2):165-173. [doi: [10.1016/j.amepre.2008.09.040](https://doi.org/10.1016/j.amepre.2008.09.040)] [Medline: [19135907](https://pubmed.ncbi.nlm.nih.gov/19135907/)]
16. Ybarra ML, Holtrop JS, Bağcı Bosi AT, Emri S. Design considerations in developing a text messaging program aimed at smoking cessation. *J Med Internet Res* 2012 Jul 24;14(4):e103 [FREE Full text] [doi: [10.2196/jmir.2061](https://doi.org/10.2196/jmir.2061)] [Medline: [22832182](https://pubmed.ncbi.nlm.nih.gov/22832182/)]
17. Hoermann S, McCabe KL, Milne DN, Calvo RA. Application of synchronous text-based dialogue systems in mental health interventions: systematic review. *J Med Internet Res* 2017 Jul 21;19(8):e267 [FREE Full text] [doi: [10.2196/jmir.7023](https://doi.org/10.2196/jmir.7023)] [Medline: [28784594](https://pubmed.ncbi.nlm.nih.gov/28784594/)]
18. Willoughby JF, Liu S. Do pictures help tell the story? An experimental test of narrative and emojis in a health text message intervention. *Comput Hum Behav* 2018 Feb;79:75-82. [doi: [10.1016/j.chb.2017.10.031](https://doi.org/10.1016/j.chb.2017.10.031)]
19. Perera AI, Thomas MG, Moore JO, Faasse K, Petrie KJ. Effect of a smartphone application incorporating personalized health-related imagery on adherence to antiretroviral therapy: a randomized clinical trial. *AIDS Patient Care STDS* 2014 Nov;28(11):579-586 [FREE Full text] [doi: [10.1089/apc.2014.0156](https://doi.org/10.1089/apc.2014.0156)] [Medline: [25290556](https://pubmed.ncbi.nlm.nih.gov/25290556/)]
20. Frøisland DH, Arsand E, Skårderud F. Improving diabetes care for young people with type 1 diabetes through visual learning on mobile phones: mixed-methods study. *J Med Internet Res* 2012 Aug 06;14(4):e111 [FREE Full text] [doi: [10.2196/jmir.2155](https://doi.org/10.2196/jmir.2155)] [Medline: [22868871](https://pubmed.ncbi.nlm.nih.gov/22868871/)]
21. Whittaker R, Maddison R, McRobbie H, Bullen C, Denny S, Dorey E, et al. A multimedia mobile phone-based youth smoking cessation intervention: findings from content development and piloting studies. *J Med Internet Res* 2008 Nov 25;10(5):e49 [FREE Full text] [doi: [10.2196/jmir.1007](https://doi.org/10.2196/jmir.1007)] [Medline: [19033148](https://pubmed.ncbi.nlm.nih.gov/19033148/)]
22. Eakin EG, Lichtenstein E, Severson HH, Stevens VJ, Vogt TM, Hollis JF. Use of tailored videos in primary care smoking cessation interventions. *Health Educ Res* 1998 Dec;13(4):519-527. [doi: [10.1093/her/13.4.519](https://doi.org/10.1093/her/13.4.519)]
23. Abrams LC, Whittaker R, Free C, Mendel Van Alstyne J, Schindler-Ruwisch JM. Developing and pretesting a text messaging program for health behavior change: recommended steps. *JMIR Mhealth Uhealth* 2015 Dec 21;3(4):e107 [FREE Full text] [doi: [10.2196/mhealth.4917](https://doi.org/10.2196/mhealth.4917)] [Medline: [26690917](https://pubmed.ncbi.nlm.nih.gov/26690917/)]
24. Attribution 4.0 International (CC BY 4.0). Creative Commons. URL: <https://creativecommons.org/licenses/by/4.0/> [accessed 2024-10-08]
25. Maar MA, Yeates K, Toth Z, Barron M, Boesch L, Hua-Stewart D, et al. Unpacking the black box: a formative research approach to the development of theory-driven, evidence-based, and culturally safe text messages in mobile health interventions. *JMIR Mhealth Uhealth* 2016 Jan 22;4(1):e10 [FREE Full text] [doi: [10.2196/mhealth.4994](https://doi.org/10.2196/mhealth.4994)] [Medline: [26800712](https://pubmed.ncbi.nlm.nih.gov/26800712/)]
26. Cole-Lewis H, Kershaw T. Text messaging as a tool for behavior change in disease prevention and management. *Epidemiol Rev* 2010;32(1):56-69 [FREE Full text] [doi: [10.1093/epirev/mxq004](https://doi.org/10.1093/epirev/mxq004)] [Medline: [20354039](https://pubmed.ncbi.nlm.nih.gov/20354039/)]
27. MacPherson MM, Cranston KD, Locke SR, Bourne JE, Jung ME. Using the behavior change wheel to develop text messages to promote diet and physical activity adherence following a diabetes prevention program. *Transl Behav Med* 2021 Aug 13;11(8):1585-1595 [FREE Full text] [doi: [10.1093/tbm/ibab058](https://doi.org/10.1093/tbm/ibab058)] [Medline: [34008852](https://pubmed.ncbi.nlm.nih.gov/34008852/)]
28. Bartlett YK, Farmer A, Rea R, French DP. Use of brief messages based on behavior change techniques to encourage medication adherence in people with type 2 diabetes: developmental studies. *J Med Internet Res* 2020 May 13;22(5):e15989 [FREE Full text] [doi: [10.2196/15989](https://doi.org/10.2196/15989)] [Medline: [32401214](https://pubmed.ncbi.nlm.nih.gov/32401214/)]
29. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. 2017 Presented at: NIPS'17; December 4-9, 2017; Long Beach, CA.

30. Devlin J, Chang MW, Lee K, Toutanova K. Bert: pre-training of deep bidirectional transformers for language understanding. arXiv Preprint posted online on October 11, 2018.
31. Radford A, Wu J, Child R, Luan D, Amodei D, Sutskever I. Language models are unsupervised multitask learners. OpenAI Blog 2019;1(8):9 [FREE Full text]
32. Brown TB, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, et al. Language models are few-shot learners. arXiv Preprint posted online on May 28, 2020 [FREE Full text]
33. Bubeck S, Chandrasekaran V, Eldan R, Gehrke J, Horvitz E, Kamar E, et al. Sparks of artificial general intelligence: early experiments with GPT-4. arXiv Preprint posted online on March 22, 2023 [FREE Full text]
34. Nori H, King N, McKinney SM, Carignan D, Horvitz E. Capabilities of GPT-4 on medical challenge problems. arXiv Preprint posted online on March 20, 2023 [FREE Full text]
35. West CG. Advances in apparent conceptual physics reasoning in GPT-4. arXiv Preprint posted online on March 29, 2023 [FREE Full text]
36. Park JS, O'Brien J, Cai CJ, Morris MR, Liang P, Bernstein MS. Generative agents: interactive simulacra of human behavior. In: Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology. 2023 Presented at: UIST '23; October 29-November 1, 2023; San Francisco, CA. [doi: [10.1145/3586183.3606763](https://doi.org/10.1145/3586183.3606763)]
37. Vaithilingam P, Zhang T, Glassman EL. Expectation vs. experience: evaluating the usability of code generation tools powered by large language models. In: Proceedings of the CHI Conference on Human Factors in Computing Systems Extended Abstracts. 2022 Presented at: CHI EA '22; April 29-May 5, 2022; New Orleans, LA. [doi: [10.1145/3491101.3519665](https://doi.org/10.1145/3491101.3519665)]
38. Anaby-Tavor A, Carmeli B, Goldbraich E, Kantor A, Kour G, Shlomov S, et al. Do not have enough data? Deep learning to the rescue!. Proc AAAI Conf Artif Intell 2020 Apr 03;34(05):7383-7390. [doi: [10.1609/aaai.v34i05.6233](https://doi.org/10.1609/aaai.v34i05.6233)]
39. Puri R, Spring R, Shoeybi M, Patwary M, Catanzaro B. Training question answering models from synthetic data. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing. 2020 Presented at: EMNLP 2020; November 16-20, 2020; Online. [doi: [10.18653/v1/2020.emnlp-main.468](https://doi.org/10.18653/v1/2020.emnlp-main.468)]
40. Kumar V, Choudhary A, Cho E. Data augmentation using pre-trained transformer models. In: Proceedings of the 2nd Workshop on Life-long Learning for Spoken Language Systems. 2020 Presented at: lifelongnlp 2020; December 7, 2020; Online.
41. Wu T, He S, Liu J, Sun S, Liu K, Han QL, et al. A brief overview of ChatGPT: the history, status quo and potential future development. IEEE CAA J Autom Sinica 2023 May;10(5):1122-1136. [doi: [10.1109/jas.2023.123618](https://doi.org/10.1109/jas.2023.123618)]
42. Liu Y, Han T, Ma S, Zhang J, Yang Y, Tian J, et al. Summary of ChatGPT-related research and perspective towards the future of large language models. Meta Radiol 2023 Sep;1(2):100017. [doi: [10.1016/j.metrad.2023.100017](https://doi.org/10.1016/j.metrad.2023.100017)]
43. Zhang Z, Gu Y, Han X, Chen S, Xiao C, Sun Z, et al. CPM-2: large-scale cost-effective pre-trained language models. AI Open 2021;2:216-224. [doi: [10.1016/j.aiopen.2021.12.003](https://doi.org/10.1016/j.aiopen.2021.12.003)]
44. Liu P, Yuan W, Fu J, Jiang Z, Hayashi H, Neubig G. Pre-train, prompt, and predict: a systematic survey of prompting methods in natural language processing. ACM Comput Surv 2023 Jan 16;55(9):1-35. [doi: [10.1145/3560815](https://doi.org/10.1145/3560815)]
45. White J, Fu Q, Hays S, Sandborn M, Olea C, Gilbert H, et al. A prompt pattern catalog to enhance prompt engineering with ChatGPT. arXiv Preprint posted online on February 21, 2023 [FREE Full text]
46. Bensoussan BE, Fleisher CS. Analysis Without Paralysis: 12 Tools to Make Better Strategic Decisions. London, UK: Pearson Education; 2012.
47. API reference. OpenAI Platform. URL: <https://platform.openai.com/docs/api-reference/> [accessed 2023-08-08]
48. FAQ. OpenAI Platform. URL: <https://platform.openai.com/docs/guides/gpt/chat-completions-vs-completions> [accessed 2023-08-08]
49. Chatterjee S, Khunti K, Davies MJ. Type 2 diabetes. Lancet 2017 Jun 03;389(10085):2239-2251. [doi: [10.1016/S0140-6736\(17\)30058-2](https://doi.org/10.1016/S0140-6736(17)30058-2)] [Medline: [28190580](https://pubmed.ncbi.nlm.nih.gov/28190580/)]
50. NCD Risk Factor Collaboration (NCD-RisC). Worldwide trends in diabetes since 1980: a pooled analysis of 751 population-based studies with 4.4 million participants. Lancet 2016 Apr 09;387(10027):1513-1530 [FREE Full text] [doi: [10.1016/S0140-6736\(16\)00618-8](https://doi.org/10.1016/S0140-6736(16)00618-8)] [Medline: [27061677](https://pubmed.ncbi.nlm.nih.gov/27061677/)]
51. Cramer JA. A systematic review of adherence with medications for diabetes. Diabetes Care 2004 May;27(5):1218-1224. [doi: [10.2337/diacare.27.5.1218](https://doi.org/10.2337/diacare.27.5.1218)] [Medline: [15111553](https://pubmed.ncbi.nlm.nih.gov/15111553/)]
52. Cramer JA, Benedict A, Muszbek N, Keskinaslan A, Khan ZM. The significance of compliance and persistence in the treatment of diabetes, hypertension and dyslipidaemia: a review. Int J Clin Pract 2008 Jan;62(1):76-87 [FREE Full text] [doi: [10.1111/j.1742-1241.2007.01630.x](https://doi.org/10.1111/j.1742-1241.2007.01630.x)] [Medline: [17983433](https://pubmed.ncbi.nlm.nih.gov/17983433/)]
53. Nieuwlaat R, Wilczynski N, Navarro T, Hobson N, Jeffery R, Keenanasseril A, et al. Interventions for enhancing medication adherence. Cochrane Database Syst Rev 2014 Nov 20;2014(11):CD000011 [FREE Full text] [doi: [10.1002/14651858.CD000011.pub4](https://doi.org/10.1002/14651858.CD000011.pub4)] [Medline: [25412402](https://pubmed.ncbi.nlm.nih.gov/25412402/)]
54. Rathbone AL, Prescott J. The use of mobile apps and SMS messaging as physical and mental health interventions: systematic review. J Med Internet Res 2017 Aug 24;19(8):e295 [FREE Full text] [doi: [10.2196/jmir.7740](https://doi.org/10.2196/jmir.7740)] [Medline: [28838887](https://pubmed.ncbi.nlm.nih.gov/28838887/)]
55. Kamal AK, Shaikh Q, Pasha O, Azam I, Islam M, Memon AA, et al. A randomized controlled behavioral intervention trial to improve medication adherence in adult stroke patients with prescription tailored Short Messaging Service

- (SMS)-SMS4Stroke study. *BMC Neurol* 2015 Oct 21;15:212 [FREE Full text] [doi: [10.1186/s12883-015-0471-5](https://doi.org/10.1186/s12883-015-0471-5)] [Medline: [26486857](https://pubmed.ncbi.nlm.nih.gov/26486857/)]
56. Patrick K, Raab F, Adams MA, Dillon L, Zabinski M, Rock CL, et al. A text message-based intervention for weight loss: randomized controlled trial. *J Med Internet Res* 2009 Jan 13;11(1):e1 [FREE Full text] [doi: [10.2196/jmir.1100](https://doi.org/10.2196/jmir.1100)] [Medline: [19141433](https://pubmed.ncbi.nlm.nih.gov/19141433/)]
 57. Finitis DJ, Pellowski JA, Johnson BT. Text message intervention designs to promote adherence to antiretroviral therapy (ART): a meta-analysis of randomized controlled trials. *PLoS One* 2014 Feb 5;9(2):e88166 [FREE Full text] [doi: [10.1371/journal.pone.0088166](https://doi.org/10.1371/journal.pone.0088166)] [Medline: [24505411](https://pubmed.ncbi.nlm.nih.gov/24505411/)]
 58. Arora S, Peters AL, Burner E, Lam CN, Menchine M. Trial to examine text message-based mHealth in emergency department patients with diabetes (TExT-MED): a randomized controlled trial. *Ann Emerg Med* 2014 Jun;63(6):745-54.e6. [doi: [10.1016/j.annemergmed.2013.10.012](https://doi.org/10.1016/j.annemergmed.2013.10.012)] [Medline: [24225332](https://pubmed.ncbi.nlm.nih.gov/24225332/)]
 59. Brath H, Morak J, Kästenbauer T, Modre-Osprian R, Strohner-Kästenbauer H, Schwarz M, et al. Mobile health (mHealth) based medication adherence measurement - a pilot trial using electronic blisters in diabetes patients. *Br J Clin Pharmacol* 2013 Sep;76 Suppl 1(Suppl 1):47-55 [FREE Full text] [doi: [10.1111/bcp.12184](https://doi.org/10.1111/bcp.12184)] [Medline: [24007452](https://pubmed.ncbi.nlm.nih.gov/24007452/)]
 60. Shetty AS, Chamukuttan S, Nanditha A, Raj RK, Ramachandran A. Reinforcement of adherence to prescription recommendations in Asian Indian diabetes patients using short message service (SMS)--a pilot study. *J Assoc Physicians India* 2011 Nov;59:711-714. [Medline: [22616337](https://pubmed.ncbi.nlm.nih.gov/22616337/)]
 61. Farmer AJ, McSharry J, Rowbotham S, McGowan L, Ricci-Cabello I, French DP. Effects of interventions promoting monitoring of medication use and brief messaging on medication adherence for people with type 2 diabetes: a systematic review of randomized trials. *Diabet Med* 2016 May;33(5):565-579. [doi: [10.1111/dme.12987](https://doi.org/10.1111/dme.12987)] [Medline: [26470750](https://pubmed.ncbi.nlm.nih.gov/26470750/)]
 62. Michie S, Abraham C, Eccles MP, Francis JJ, Hardeman W, Johnston M. Strengthening evaluation and implementation by specifying components of behaviour change interventions: a study protocol. *Implement Sci* 2011 Feb 07;6:10 [FREE Full text] [doi: [10.1186/1748-5908-6-10](https://doi.org/10.1186/1748-5908-6-10)] [Medline: [21299860](https://pubmed.ncbi.nlm.nih.gov/21299860/)]
 63. Gellman MD. Behavioral medicine. In: Gellman MD, editor. *Encyclopedia of Behavioral Medicine*. Cham, Switzerland: Springer; 2020.
 64. Michie S, Richardson M, Johnston M, Abraham C, Francis J, Hardeman W, et al. The behavior change technique taxonomy (v1) of 93 hierarchically clustered techniques: building an international consensus for the reporting of behavior change interventions. *Ann Behav Med* 2013 Aug;46(1):81-95 [FREE Full text] [doi: [10.1007/s12160-013-9486-6](https://doi.org/10.1007/s12160-013-9486-6)] [Medline: [23512568](https://pubmed.ncbi.nlm.nih.gov/23512568/)]
 65. Dobson R, Whittaker R, Pfaeffli Dale L, Maddison R. The effectiveness of text message-based self-management interventions for poorly-controlled diabetes: a systematic review. *Digit Health* 2017 Nov 09;3:2055207617740315 [FREE Full text] [doi: [10.1177/2055207617740315](https://doi.org/10.1177/2055207617740315)] [Medline: [29942620](https://pubmed.ncbi.nlm.nih.gov/29942620/)]
 66. Kebede MM, Liedtke TP, Möllers T, Pischke CR. Characterizing active ingredients of eHealth interventions targeting persons with poorly controlled type 2 diabetes mellitus using the behavior change techniques taxonomy: scoping review. *J Med Internet Res* 2017 Oct 12;19(10):e348 [FREE Full text] [doi: [10.2196/jmir.7135](https://doi.org/10.2196/jmir.7135)] [Medline: [29025693](https://pubmed.ncbi.nlm.nih.gov/29025693/)]
 67. Webb TL, Joseph J, Yardley L, Michie S. Using the internet to promote health behavior change: a systematic review and meta-analysis of the impact of theoretical basis, use of behavior change techniques, and mode of delivery on efficacy. *J Med Internet Res* 2010 Feb 17;12(1):e4 [FREE Full text] [doi: [10.2196/jmir.1376](https://doi.org/10.2196/jmir.1376)] [Medline: [20164043](https://pubmed.ncbi.nlm.nih.gov/20164043/)]
 68. Long H, Bartlett YK, Farmer AJ, French DP. Identifying brief message content for interventions delivered via mobile devices to improve medication adherence in people with type 2 diabetes mellitus: a rapid systematic review. *J Med Internet Res* 2019 Jan 09;21(1):e10421 [FREE Full text] [doi: [10.2196/10421](https://doi.org/10.2196/10421)] [Medline: [30626562](https://pubmed.ncbi.nlm.nih.gov/30626562/)]
 69. Bommakanti KK, Smith LL, Liu L, Do D, Cuevas-Mota J, Collins K, et al. Requiring smartphone ownership for mHealth interventions: who could be left out? *BMC Public Health* 2020 Jan 20;20(1):81 [FREE Full text] [doi: [10.1186/s12889-019-7892-9](https://doi.org/10.1186/s12889-019-7892-9)] [Medline: [31959145](https://pubmed.ncbi.nlm.nih.gov/31959145/)]
 70. Graesser AC, McNamara DS, Kulikowich JM. Coh-matrix: providing multilevel analyses of text characteristics. *Educ Res* 2011 Jun 01;40(5):223-234. [doi: [10.3102/0013189x11413260](https://doi.org/10.3102/0013189x11413260)]
 71. Nelson J, Perfetti C, Liben D, Liben M. Measures of text difficulty: testing their predictive value for grade levels and student performance. *Student Achievement Partners*. 2012. URL: <https://achievethecore.org/page/1196/measures-of-text-difficulty-testing-their-predictive-value-for-grade-levels-and-student-performance> [accessed 2023-08-18]
 72. Valueva EA, Danilevskaya NM, Lapteva EM, Ushakov DV. Phenomenon of secular iq gains: the analysis of children's fiction. *Psikhologicheskii Zhurnal* 2017;38(5):18-26 [FREE Full text] [doi: [10.7868/S0205959217050026](https://doi.org/10.7868/S0205959217050026)]
 73. Frantz RS, Starr LE, Bailey AL. Syntactic complexity as an aspect of text complexity. *Educ Res* 2015 Oct 01;44(7):387-393. [doi: [10.3102/0013189x15603980](https://doi.org/10.3102/0013189x15603980)]
 74. Text messaging in healthcare research toolkit. Center for Research in Implementation Science and Prevention (CRISP), University of Colorado School of Medicine. 2017. URL: https://www.careinnovations.org/wp-content/uploads/2017/11/Text_Messaging_in_Healthcare_Research_Toolkit_2.pdf [accessed 2023-08-18]
 75. Ranney ML, Choo EK, Cunningham RM, Spirito A, Thorsen M, Mello MJ, et al. Acceptability, language, and structure of text message-based behavioral interventions for high-risk adolescent females: a qualitative study. *J Adolesc Health* 2014 Jul;55(1):33-40 [FREE Full text] [doi: [10.1016/j.jadohealth.2013.12.017](https://doi.org/10.1016/j.jadohealth.2013.12.017)] [Medline: [24559973](https://pubmed.ncbi.nlm.nih.gov/24559973/)]

76. Thoppilan R, De Freitas D, Hall J, Shazeer N, Kulshreshtha A, Cheng HT, et al. Lamda: language models for dialog applications. arXiv Preprint posted online on January 20, 2022 [[FREE Full text](#)]
77. Mukherjee S, Mitra A, Jawahar G, Agarwal S, Palangi H, Awadallah A. Orca: progressive learning from complex explanation traces of gpt-4. arXiv Preprint posted online on June 5, 2023 [[FREE Full text](#)]
78. Touvron H, Martin L, Stone K, Albert P, Almahairi A, Babaei Y, et al. Llama 2: open foundation and fine-tuned chat models. arXiv Preprint posted online on July 18, 2023 [[FREE Full text](#)]
79. Vogels EA. A majority of Americans have heard of ChatGPT, but few have tried it themselves. Pew Research Center. 2023 May 24. URL: <https://www.pewresearch.org/short-reads/2023/05/24/a-majority-of-americans-have-heard-of-chatgpt-but-few-have-tried-it-themselves/> [accessed 2023-12-09]
80. Hu K. ChatGPT sets record for fastest-growing user base - analyst note. Reuters. 2023 Feb 2. URL: <https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/> [accessed 2023-12-09]
81. Yu Y, Zhuang Y, Zhang J, Meng Y, Ratner A, Krishna R, et al. Large language model as attributed training data generator: a tale of diversity and bias. arXiv Preprint posted online on June 28, 2023 [[FREE Full text](#)]
82. Code of Federal Regulations. US Department of Health and Human Services. URL: <https://www.ecfr.gov/on/2018-07-19/title-45/subtitle-A/subchapter-A/part-46> [accessed 2024-09-09]
83. google-research / bert. GitHub. URL: <https://github.com/google-research/bert> [accessed 2023-08-08]
84. BERT base model (uncased). Hugging Face. URL: <https://huggingface.co/bert-base-uncased> [accessed 2023-08-27]
85. Abdi H, Williams LJ. Principal component analysis. WIREs Comput Stat 2010 Jul 15;2(4):433-459. [doi: [10.1002/wics.101](https://doi.org/10.1002/wics.101)]
86. New and improved embedding model. OpenAI. URL: <https://openai.com/blog/new-and-improved-embedding-model> [accessed 2023-08-27]
87. Embeddings. OpenAI Platform. URL: <https://platform.openai.com/docs/guides/embeddings> [accessed 2023-08-08]
88. van der Maaten L, Hinton G. Visualizing data using t-SNE. J Mach Learn Res 2008;9:2579-2605.
89. Harrison RM, Dereventsov A, Bibin A. Zero-shot recommendations with pre-trained large language models for multimodal nudging. In: Proceedings of the IEEE International Conference on Data Mining Workshops. 2023 Presented at: ICDMW 2023; December 1-4, 2023; Shanghai, China. [doi: [10.1109/icdmw60847.2023.00195](https://doi.org/10.1109/icdmw60847.2023.00195)]
90. Doğru OC, Webb TL, Norman P. Can behavior change techniques be delivered via short text messages? Transl Behav Med 2022 Nov 16;12(10):979-986. [doi: [10.1093/tbm/ibac058](https://doi.org/10.1093/tbm/ibac058)] [Medline: [36190350](https://pubmed.ncbi.nlm.nih.gov/36190350/)]
91. Lu Y, Bartolo M, Moore A, Riedel S, Stenetorp P. Fantastically ordered prompts and where to find them: overcoming few-shot prompt order sensitivity. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics. 2022 Presented at: ACL 2022; May 22-27, 2022; Dublin, Ireland. [doi: [10.18653/v1/2022.acl-long.556](https://doi.org/10.18653/v1/2022.acl-long.556)]
92. Zhao TZ, Wallace E, Feng S, Klein D, Singh S. Calibrate before use: improving few-shot performance of language models. arXiv Preprint posted online on February 19, 2021 [[FREE Full text](#)]
93. Ali Khan A, Badshah S, Liang P, Waseem M, Khan B, Ahmad A, et al. Ethics of AI: a systematic literature review of principles and challenges. In: Proceedings of the 26th International Conference on Evaluation and Assessment in Software Engineering. 2022 Presented at: EASE '22; June 13-15, 2022; Gothenburg, Sweden. [doi: [10.1145/3530019.3531329](https://doi.org/10.1145/3530019.3531329)]
94. Bostrom N, Yudkowsky E. The ethics of artificial intelligence. In: Artificial Intelligence Safety and Security. Boca Raton, FL: Chapman and Hall/CRC; 2018.
95. El-Mhamdi EM, Farhadkhanis S, Guerraoui R, Gupta N, Hoang LN, Pinot R, et al. On the impossible safety of large AI models. arXiv Preprint posted online on September 30, 2022 [[FREE Full text](#)]
96. Wolf Y, Wies N, Avnery O, Levine Y, Shashua A. Fundamental limitations of alignment in large language models. arXiv Preprint posted online on April 19, 2023 [[FREE Full text](#)]
97. Nelligan RK, Hinman RS, Atkins L, Bennell KL. A short message service intervention to support adherence to home-based strengthening exercise for people with knee osteoarthritis: intervention design applying the behavior change wheel. JMIR Mhealth Uhealth 2019 Oct 18;7(10):e14619 [[FREE Full text](#)] [doi: [10.2196/14619](https://doi.org/10.2196/14619)] [Medline: [31628786](https://pubmed.ncbi.nlm.nih.gov/31628786/)]
98. Green SM, French DP, Hall LH, Bartlett YK, Rousseau N, Raine E, ROSETA Investigators, et al. Codevelopment of a text messaging intervention to support adherence to adjuvant endocrine therapy in women with breast cancer: mixed methods approach. J Med Internet Res 2023 May 24;25:e38073 [[FREE Full text](#)] [doi: [10.2196/38073](https://doi.org/10.2196/38073)] [Medline: [37223964](https://pubmed.ncbi.nlm.nih.gov/37223964/)]
99. Park LG, Howie-Esquivel J, Chung ML, Dracup K. A text messaging intervention to promote medication adherence for patients with coronary heart disease: a randomized controlled trial. Patient Educ Couns 2014 Feb;94(2):261-268. [doi: [10.1016/j.pec.2013.10.027](https://doi.org/10.1016/j.pec.2013.10.027)] [Medline: [24321403](https://pubmed.ncbi.nlm.nih.gov/24321403/)]
100. Arambepola C, Ricci-Cabello I, Manikavasagam P, Roberts N, French DP, Farmer A. The impact of automated brief messages promoting lifestyle changes delivered via mobile devices to people with type 2 diabetes: a systematic literature review and meta-analysis of controlled trials. J Med Internet Res 2016 Apr 19;18(4):e86 [[FREE Full text](#)] [doi: [10.2196/jmir.5425](https://doi.org/10.2196/jmir.5425)] [Medline: [27095386](https://pubmed.ncbi.nlm.nih.gov/27095386/)]
101. Orr JA, King RJ. Mobile phone SMS messages can enhance healthy behaviour: a meta-analysis of randomised controlled trials. Health Psychol Rev 2015;9(4):397-416. [doi: [10.1080/17437199.2015.1022847](https://doi.org/10.1080/17437199.2015.1022847)] [Medline: [25739668](https://pubmed.ncbi.nlm.nih.gov/25739668/)]
102. Wang J, Shi E, Yu S, Wu Z, Ma C, Dai H, et al. Prompt engineering for healthcare: methodologies and applications. arXiv Preprint posted online on April 28, 2023 [[FREE Full text](#)]

103. Reynolds L, McDonell K. Prompt programming for large language models: beyond the few-shot paradigm. In: Proceedings of the Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems. 2021 Presented at: CHI EA '21; May 8-13, 2021; Yokohama, Japan. [doi: [10.1145/3411763.3451760](https://doi.org/10.1145/3411763.3451760)]

Abbreviations

AI: artificial intelligence
API: application programming interface
BCT: behavior change technique
GSM-7: Global System for Mobile Communications
LLM: large language model
mHealth: mobile health
PCA: principal component analysis
t-SNE: t-distributed stochastic neighbor embedding

Edited by K El Emam, B Malin; submitted 20.09.23; peer-reviewed by A Martins, C Rios-Bedoya, K Andreadis, R Odabashian; comments to author 04.12.23; revised version received 10.12.23; accepted 13.06.24; published 15.10.24.

Please cite as:

Harrison RM, Lapteva E, Bibin A

Behavioral Nudging With Generative AI for Content Development in SMS Health Care Interventions: Case Study

JMIR AI 2024;3:e52974

URL: <https://ai.jmir.org/2024/1/e52974>

doi: [10.2196/52974](https://doi.org/10.2196/52974)

PMID: [39405108](https://pubmed.ncbi.nlm.nih.gov/39405108/)

©Rachel M Harrison, Ekaterina Lapteva, Anton Bibin. Originally published in JMIR AI (<https://ai.jmir.org>), 15.10.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR AI, is properly cited. The complete bibliographic information, a link to the original publication on <https://www.ai.jmir.org/>, as well as this copyright and license information must be included.

Original Paper

Identifying Links Between Productivity and Biobehavioral Rhythms Modeled From Multimodal Sensor Streams: Exploratory Quantitative Study

Runze Yan¹, PhD; Xinwen Liu², MS; Janine M Dutcher², PhD; Michael J Tumminia³, PhD; Daniella Villalba², PhD; Sheldon Cohen², PhD; John D Creswell², PhD; Kasey Creswell², PhD; Jennifer Mankoff⁴, PhD; Anind K Dey⁴, PhD; Afsaneh Doryab¹, PhD

¹University of Virginia, Charlottesville, VA, United States

²Carnegie Mellon University, Pittsburgh, PA, United States

³University of Pittsburgh, Pittsburgh, PA, United States

⁴University of Washington, Seattle, WA, United States

Corresponding Author:

Afsaneh Doryab, PhD
University of Virginia
351 McCormick Road
Charlottesville, VA, 22904
United States
Phone: 1 4342435823
Email: ad4ks@virginia.edu

Abstract

Background: Biobehavioral rhythms are biological, behavioral, and psychosocial processes with repeating cycles. Abnormal rhythms have been linked to various health issues, such as sleep disorders, obesity, and depression.

Objective: This study aims to identify links between productivity and biobehavioral rhythms modeled from passively collected mobile data streams.

Methods: In this study, we used a multimodal mobile sensing data set consisting of data collected from smartphones and Fitbits worn by 188 college students over a continuous period of 16 weeks. The participants reported their self-evaluated daily productivity score (ranging from 0 to 4) during weeks 1, 6, and 15. To analyze the data, we modeled cyclic human behavior patterns based on multimodal mobile sensing data gathered during weeks 1, 6, 15, and the adjacent weeks. Our methodology resulted in the creation of a rhythm model for each sensor feature. Additionally, we developed a correlation-based approach to identify connections between rhythm stability and high or low productivity levels.

Results: Differences exist in the biobehavioral rhythms of high- and low-productivity students, with those demonstrating greater rhythm stability also exhibiting higher productivity levels. Notably, a negative correlation ($C=-0.16$) was observed between productivity and the SE of the phase for the 24-hour period during week 1, with a higher SE indicative of lower rhythm stability.

Conclusions: Modeling biobehavioral rhythms has the potential to quantify and forecast productivity. The findings have implications for building novel cyber-human systems that align with human beings' biobehavioral rhythms to improve health, well-being, and work performance.

(JMIR AI 2024;3:e47194) doi:[10.2196/47194](https://doi.org/10.2196/47194)

KEYWORDS

biobehavioral rhythms; productivity; computational modeling; mobile sensing; mobile phone

Introduction

Background

Biobehavioral rhythms—repeated cycles of biological, behavioral, and psychological events—are indicative of different life and health outcomes [1]. Chronobiology, which examines periodic phenomena in living organisms, has demonstrated the impact of circadian disruptions on people's lives, including physical and mental health as well as safety and work performance in shift workers [2-6]. However, research in chronobiology has primarily been conducted via manual observations and subjective reports often restricted over a small period of time. Advances in mobile and wearable devices provide the possibility of automatic and rigorous collection of longitudinal biobehavioral data from people's personal devices [7-9]. This longitudinal fine-grained data collected on a daily basis have the potential to reveal micro- and macrolevel patterns related to different biobehavioral outcomes.

In this study, we examine the relationship between cyclical human behaviors and work efficiency using data from mobile sensors. This analysis is based on data collected from the smartphones and Fitbits of 166 college students, encompassing patterns such as activity, communication, and sleep. Our main objective is to determine variations in biobehavioral rhythms across students of varying productivity levels and identify particular rhythm traits associated with productivity.

Related Work

Modeling Biobehavioral Rhythms

Research in chronobiology that examines periodic phenomena in living organisms is relatively mature, and existing studies have confirmed that exploring human rhythms is an effective way to diagnose and treat many illnesses such as cancer, cardiovascular disease, and mental health problems [10-12]. For example, patients with depression, those with bipolar disorder, and those with schizophrenia usually exhibit irregular changes in circadian rhythm, and adjusting the circadian rhythm is an efficient auxiliary method for treating these conditions [13-15]. Disruption in biological rhythms is also caused by changing lifestyles and environmental conditions such as travel across time zones and shift work [16]. Night shift and morning shift workers may be especially at risk of committing errors and having accidents [17].

A few studies have used smartphone technology to track circadian patterns. For example, Abdullah et al [18] used patterns of phone usage to identify chronotypes of students (early birds or night owls). Murnane et al [19] aggregated mobile app usage features by body clock time and analyzed the correlation between circadian rhythms in app usage and alertness level. Doryab et al [1] demonstrated modeling of rhythms using data from Fitbit devices in patients with cancer and showed that disruption in circadian rhythms predicts readmission in patients with cancer undergoing treatment. Yan et al [7] further developed a computational framework for modeling biobehavioral rhythms from multimodal sensor streams. While our work leverages this framework to model biobehavioral rhythms, we advance research in this domain by developing

and applying algorithms to observe and measure changes in multimodal biobehavioral rhythms across different periods and between people with different productivity levels.

Productivity Assessment

Traditional productivity assessment approaches are typically subjective, static evaluations administered as self-report surveys, manager assessments, observations, or ability tests. Some studies have used multitasking and interruptions, for example, checking emails [20] and mental and physical fatigue as proxies for productivity in workers and officers [21-24]. For example, Gloria et al [20] tracked and analyzed email usage in affecting workplace productivity and stress. Aryal et al [25] conducted a simulated construction task for monitoring physical fatigue by measuring changes in heart rate, skin temperature, and brain signals. The study showed a direct relationship between physical fatigue and heart rate metrics such as heart rate, heart rate variability, and percentage of heart rate.

Recent studies on workplace productivity have used mobile, wearable, and environmental sensors to track individuals' behavior and environmental conditions to assess workers' job performance. For example, background noise, light, temperature, and air quality have been shown as the 4 external factors affecting productivity [26-29]. In a study by Mirjafari et al [30], the analysis of phone usage, location, activity, sleep, and time allocation of 554 participants indicated that the regularity of behaviors distinguishes high and low performance. van Vugt et al [31] suggested that eye-tracking could be used to measure productivity. The hypothesis was that if the eyes of a person remained at certain locations on the computer screen, they were focused and thus productive. However, this theory has yet to be evaluated in practice. In addition to external factors, research studies have investigated the impact of internal factors and cues in measuring productivity. For example, Das Swain et al [32] demonstrated that static intrinsic personality can explain workplace performance using data from 603 information workers.

Our research is unique in measuring and assessing productivity by leveraging cyclic biobehavioral patterns from passive data streams to assess productivity. Our work is also the first to measure daily productivity from multimodal mobile and wearable data in college students.

Methods

Data Collection

We use a data set of smartphone and Fitbit logs collected from 188 students at an American university over the course of 1 semester. The data were collected as part of an extensive study on students' health and well-being. All participants were first-year students, with their demographic details presented in Table 1.

The AWARE data collection app [33] and Fitbit were used for the collection of audio, Bluetooth, Wi-Fi, location, phone usage, calls, calories, sleep, and steps. AWARE is an open-source data collection framework that works both on Android and iOS platforms. All participants used their smartphones, and this study's team provided a Fitbit Flex 2 to collect data. Students'

productivity assessments were collected via an evening survey during weeks 1, 6 (midsemester), and 15 (last week of classes) of the semesters to avoid overburdening participants. The assessment question included a single question: “How productive did you feel today?” The possible responses ranged from 0 (not productive at all) to 4 (extremely productive). The mean and SD of self-evaluated productivity scores were consistent for different sexes and major groups with no significant difference: female (mean 1.65, SD 0.92), male (mean 1.80, SD 0.97), engineering (mean 1.71, SD 0.96), business (mean 1.70, SD 0.99), science (mean 1.69, SD 0.94), art (mean

1.76, SD 0.95), humanities (mean 1.68, SD 0.97), and undecided (mean 1.67, SD 0.87).

Of the initial 188 first-year students, 166 produced subjective assessments of their respective daily productivity. The response rate fluctuated over the 3 weeks, with some students not completing the surveys. The data set included 488 total observations, represented as participant-week pairs. During the introductory meeting, students were briefed about this study’s objectives. This study’s goals were transparently communicated without any deceit or exclusion.

Table 1. Demographic distribution of this study’s samples: a total of 188 first-year university students were enlisted as participants for this research.

Category and subcategory	Participants, n (%)
Sex	
Male	111 (59)
Female	77 (41)
Race	
Asian	107 (57)
Black	9 (5)
Hispanic	17 (9)
White	64 (34)
Major	
Engineering	79 (42)
Art	30 (16)
Business	24 (13)
Science	23 (12)
Humanities	8 (4)

Data Processing

Measuring Productivity Levels

As mentioned previously, while sensor data were collected continuously for 16 weeks, self-reported productivity (by study design) was only collected in weeks 1, 6, and 15. We used productivity scores (0-4) to categorize participants into high and low-productivity groups. These categories were used as ground truth labels in the later analysis of the relationship between rhythms and productivity. To identify the cutoff threshold, we calculated the mean and median of the daily productivity scores for all participants across all 3 weeks. The mean of 1.89 (SD 0.94) and a median of 2 (IQR 1) indicated a normal distribution across scores (verified by the Shapiro-Wilk test, $P=.12$). Therefore, we used 2 as the threshold for categorizing productivity, with scores less than 2 indicating low productivity and scores equal to or above 2 indicating high productivity. Figure 1 shows the distribution of the mean and variance of daily productivity scores within each week. The mean productivity has decreased in week 6 compared with week 1. Since week 6 is the midterm, a high workload and pressure may make some students work more productively, but the

pressure and stress may have the opposite effect on others. The IQR of the mean of low productivity is wider than in week 1. The mean and 75th percentile of variance are all less than one, which is also the interval between the survey’s productivity options. This indicates that the participants’ answers are relatively stable each week. We, therefore, average the productivity scores of all days in each week (including both weekdays and weekends) as the weekly productivity score with the same threshold to categorize each participant’s week average into high or low productivity.

In addition to labeling each participant’s weekly data as high or low productivity, we also need to further categorize participants into high or low productivity to evaluate our rhythm similarity methods described in the Methods section. We analyze the combination of high and low productivity weeks for all participants as shown in Table 2. We observe the number of participants in different combinations is imbalanced and does not create large enough groups for analyzing each combination separately. We therefore categorize participants into high and low productivity groups, where students with at least 2 weeks of high productivity rates are categorized as high productivity and the rest are placed into the low-productivity group.

Figure 1. If the mean of 1 week’s daily productivity is above 2 (SD 0.21), the week will be labeled high productivity; otherwise, the week will be labeled low productivity. Gray represents the mean and variance that come from weeks with high productivity, and orange represents the mean and variance that come from weeks with low productivity. The medians of variance are all less than 0.5, and the 75 percentiles are within 1 no matter what productivity the weeks have. The difference in productivity scores between the 2 adjacent options in the productivity survey is 1, so the low variance indicates that most participants will keep the same productivity level during the whole week. The medians of the mean of both high and low productivity are very close, but there are more small mean values in week 6 for low productivity and more large mean values in week 1 for high productivity.

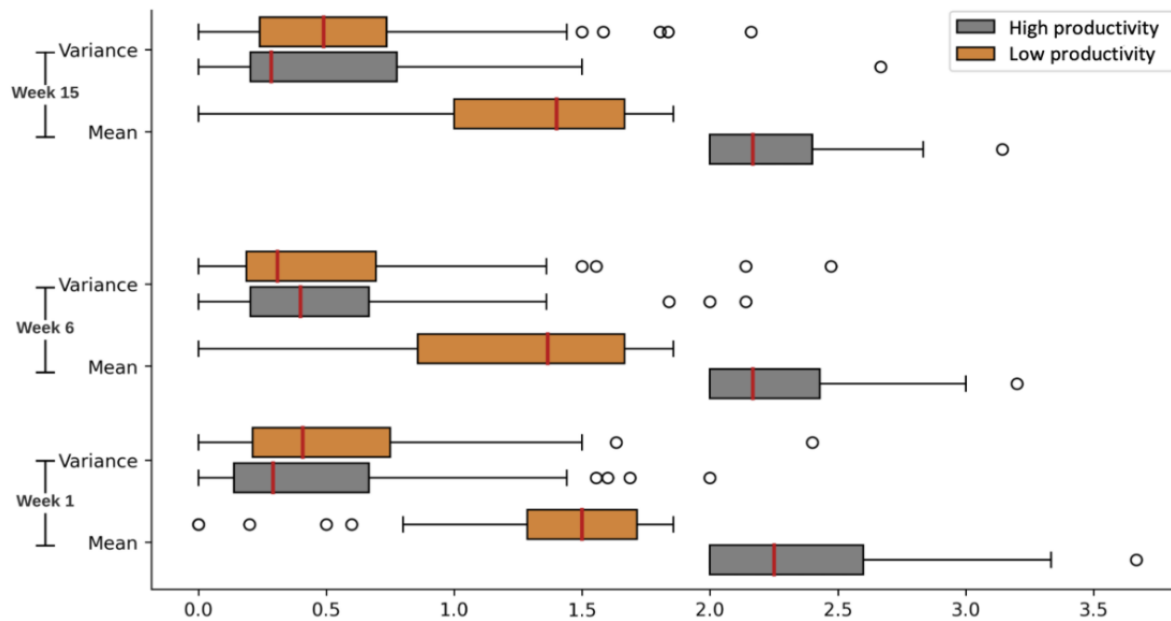


Table 2. Participant productivity^a.

Week 1	Week 6	Week 15	Participants, n
High productivity group			
High	High	High	13
High	High	Low	14
High	Low	High	12
Low	High	High	10
Low productivity group			
High	Low	Low	29
Low	High	Low	4
Low	Low	High	17
Low	Low	Low	62

^aThe middle column lists all combinations of weekly productivity levels, and the right column shows the number of participants for each combination. Many participants were inefficient for all 3 weeks. Participants were more likely to achieve high productivity in week 1 and had the most difficulty achieving high productivity in week 6. Moreover, we aggregated the 8 combinations into 2 groups. Participants with at least 2 highly productive weeks were assigned to the high-productivity group; otherwise, they were assigned to the low-productivity group.

Feature Extraction

We extracted features in 2 processing layers. First, we aggregated the raw sensor data into more meaningful behavioral features to capture students’ social interaction, physical activity, sleep, and academic life. The raw sensor data we collect are just a series of numbers without providing much information. For example, screen data are a time series of values from 0 to 3 (eg, 0121023...), which does not provide any helpful information, but we can process this time series to extract more meaningful information about how often the user has been interacting with the phone. We then divided each data stream into hourly

intervals and extracted behavioral features in each interval following the descriptions documented by Doryab et al [34]. Typical features included statistical measures such as minimum, maximum, mean, SD, length of the status in the hour, and more complex behavioral features such as movement patterns and type and duration of activities. Example features are shown in Table 3. Finally, we modeled the cyclic pattern of each behavioral feature using Cosinor, which provided a set of parameters that describe the cyclic pattern. This process and the list of rhythm parameters are detailed in the following section.

Table 3. Examples of sensor features.

Device and sensor	Extracted feature
Smartphone	
Audio	Percentage of time with voice, noise, or silence; minimum, maximum, mean, or SD of voice energy
Bluetooth	Mean or total number of Bluetooth scans
Wi-Fi	Number of unique Wi-Fi hotspots detected
Location	Location variance; percentage of time staying at home; number of visits; time spent at green areas, athletic areas, academic areas, or outside campus
Phone usage	Minutes interacting with phone; minimum, maximum, mean, or SD length of interaction periods
Fitbit	
Sleep	Minutes asleep, awake, or restless; minimum, maximum, or mean length of asleep, awake, or restless periods
Steps	Total number of steps; minimum, maximum, mean, and total length of active or sedentary periods
Calories	Minimum, maximum, mean, or total calories burned; minimum, maximum, mean, or total decrease in 5-minute calories burned

Handling Missing Values

As data sets collected in the wild are expected to include noise and missing data, we developed strategies to handle missing data. The missing values were filled separately for different participants and weeks using the local moving average commonly used in time series. For example, if the hourly values of location variance were missing at 2 PM and 3 PM on day 1 of week 1 for participant A, then we imputed the values as follows: $v_{2pm} = v_{1pm} + (v_{4pm} - v_{1pm}) / (4 - 1)$ and $v_{3pm} = v_{1pm} + 2 \times (v_{4pm} - v_{1pm}) / (4 - 1)$. Moving average is the most suitable interpolation method for rhythm modeling. Other methods such as multiple interpolations and Expectation-Maximization estimation introduce cross-correlation between features, and regression estimation and k-nearest neighbor increase auto-correlation of a single sensor feature [35,36]. However, the moving average method is sensitive to the number of continuous missing data. If the missing block is large, the moving average will introduce high noise and bias, and the data may need to be removed instead of imputed. We, therefore, calculated the average length of continuous missing hour blocks to decide the minimum threshold for removing data. The average missing block was 1.7 (SD 0.41) data points in sensor streams with less than 20% missing values. We, therefore, imputed the behavioral feature streams with less than 20% missing values and discarded the rest.

After cleaning the data, we ended up with a data set that included 101 sensor features related to location, calories, steps, and sleep. The amount of weekly data we have for each feature changes because some data from participants was removed during our missing handling process. As an example, location features have around 50 observations for week 1 and 15 and 22 observations for week 6; calories and steps features have around 110

observations for weeks 1 and 6 and 80 observations for week 15.

Modeling Bibehavioral Rhythms

To model rhythms from longitudinal bibehavioral data collected in the wild, we used the Cosinor method introduced by Halberg [37]. The Cosinor method forms a linear combination of cosine curves with known frequencies to fit cyclic time-series rhythm data and calculates rhythm parameters using least square regression [38]. The Cosinor function can take multiple periods as input parameters and use those to generate a cyclic model of provided time series data. The generated model includes a series of parameters that characterize the cyclic behavior in the data stream. **Textbox 1** details the parameters, and **Figure 2** [39] visually represents them. The Cosinor method is mathematically expressed by Fernández et al [40] as:



where y_i is the observation at time t_i ; M is the Midline Estimating Statistic of Rhythm (MESOR); t_i is the sampling time; n is the number of input periods; A_c , T_c , and ϕ_c represent the amplitude (Amp), period, and acrophase (PHI), respectively; and ϵ is the error. Cosinor also outputs the SE for MESOR, Amp, and PHI, respectively.

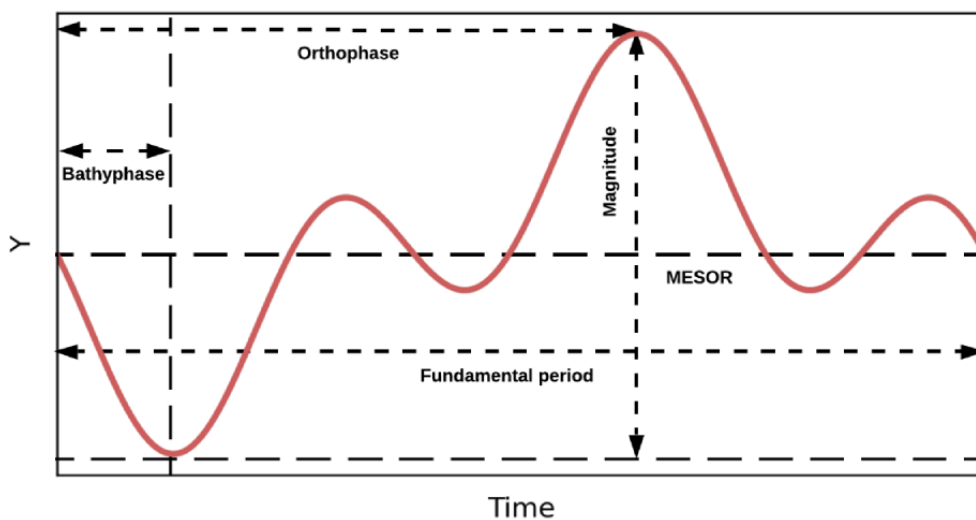
We used Cosinor to build personal cyclic models per student per sensor stream in weeks 1, 6, 15, and the weeks adjacent to them (eg, for week 6, we use sensor data from weeks 5, 6, and 7 to build Cosinor models). We then used rhythm parameters generated by those models in the correlation analysis. We assumed all participants had normal daily rhythms and used the input periods of 8, 12, and 24 hours in the Cosinor. The 8, 12, and 24 hours reflect nocturnal, diurnal, and circadian duration, respectively.

Textbox 1. Definitions of rhythm parameters output from the Cosinor model [41].

Rhythm parameters and their definition

- Fundamental period: the fundamental period is the least common multiple (LCM) of all individual periods. We use 8-, 12-, and 24-hour periods in our modeling approach.
- MESOR: estimating the midline of the rhythm curves.
- Amplitude (Amp): half the difference between the maximum and the minimum of the best-fitted curve in an individual period.
- Acrophase (PHI): lag from a defined reference time point to the maximum point within an individual period.
- Magnitude: half the difference between the maximum and the minimum of the best-fitted curve in the fundamental period.
- Bathyphase: lag from a defined reference time point to the minimum point within an individual period.
- Orthophase: lag from a defined reference time point to the maximum point within the fundamental period.
- P value (P): P value indicates the significance level of the model fitted by an individual period.
- Percent rhythm (PR): percent rhythm is the coefficient of determination (R^2) for the model using an individual period.
- Integrated P value (IP): the integrated P value indicates the significance level (P value) of the model fitted by the fundamental period.
- Integrated percent rhythm (IPR): integrated percent rhythm is the (R^2) for the model using the fundamental period.

Figure 2. The cyclic wave is formed by fundamental parameters described in Table 3 (adapted from Cornelissen [7]). MESOR: Midline Estimating Statistic of Rhythm.



Measuring the Relationship Between Rhythms and Productivity

We adopted the Pearson correlation analysis to identify relationships between rhythms and productivity across time windows (here weeks). Such a relationship, however, is multidimensional, involving multiple sensors, features, and rhythm parameters. To quantify this multidimensional relationship, we developed a 2-step method. First, we calculated the correlation coefficient between each rhythm parameter and productivity score to understand how rhythm parameters correlate with productivity and whether the correlation is consistent across weeks. To account for the varied scales of productivity and rhythm parameters, we initially applied minimum-maximum normalization to both the productivity scores and each rhythm parameter. Following this, we computed the Pearson correlation coefficient and determined its significance using a 2-tailed P value test. The first step resulted in 1 correlation coefficient and 1 P value per behavioral feature,

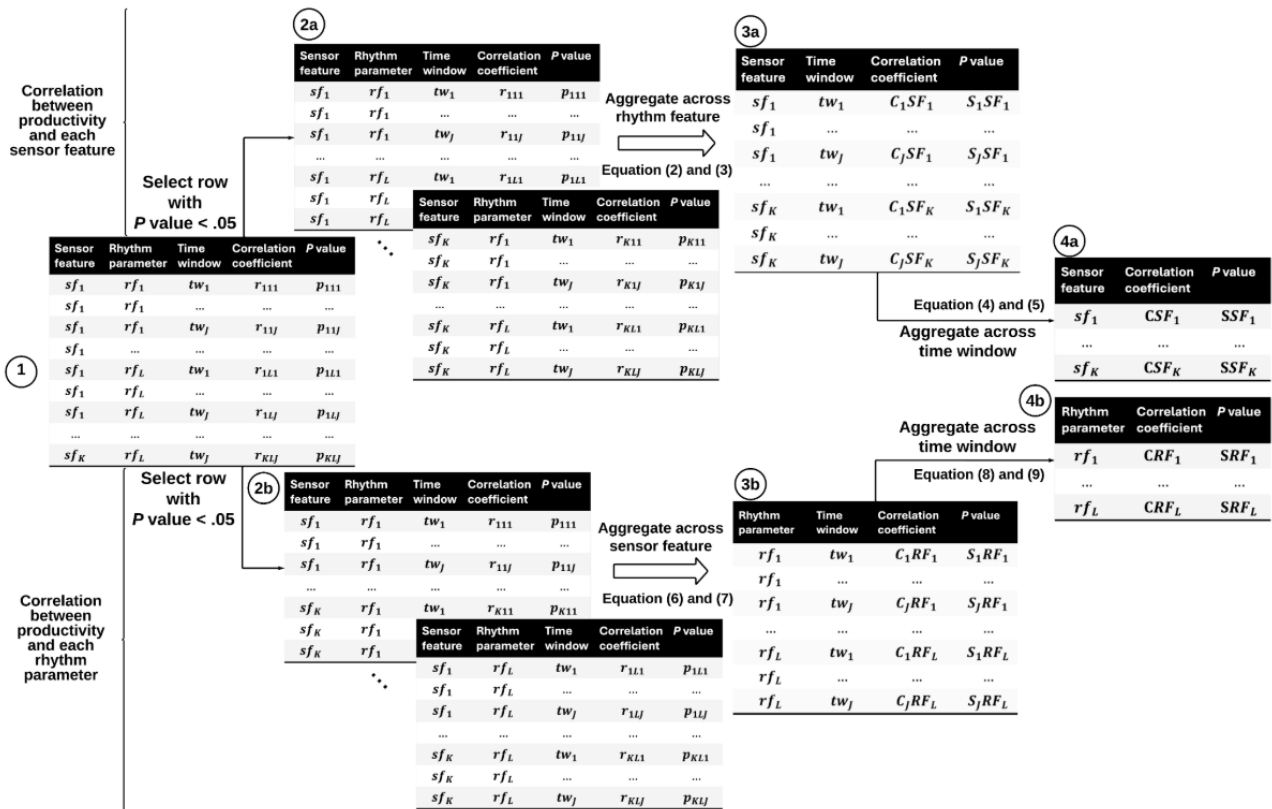
per rhythm parameter, and per time window (week) as shown in Figure 3 (step 1). The correlation coefficient indicates how closely the rhythm parameter and productivity score are related, and whether they move together or in opposite ways. The P value helps us understand if this relationship is significant or merely coincidental.

Next, as presented in Figure 3, we adopt the Fisher method to combine the correlation coefficient and its significance (P value) of every combination of behavioral feature—rhythm parameter—week. The Fisher method is a widely used meta-analysis technique used for combining the results from several independence tests [42,43]. These combinations provide information about productivity-related variations of the rhythms for each behavioral feature per week (2a in Figure 3) and productivity-related variations of each rhythm parameter per week (2b in Figure 3) regardless of behavior. While the correlation coefficient represents the strength and direction of the relationship, its significance reflects the reliability and

generalizability of the relationship. We, therefore, aggregated significant correlation coefficients for all rhythm parameters per behavioral sensor feature (2a) as well as aggregated significant correlation coefficients for all sensor features per rhythm parameter per week (2b). In step 3 (3a and 3b), we further combined correlation coefficients and significance scores across all 3 weeks. The final step (4) summarizes the correlation (and significance) values into 1 final score for each sensor

feature (4a) and for each rhythm feature (4b). The calculation process is detailed in the [Multimedia Appendices 1 and 2](#). Since the number of observations is different for different rhythm parameters, behavioral sensor features, and weeks due to missing values, this analysis was only performed on the correlations with more than 28 observations, which is the median value in our data set.

Figure 3. The pipeline to aggregate the correlation for a multidimensional dataset with K sensor features, L rhythm parameters, and J time windows. The pipeline can output the correlation between productivity and a single sensor, and the correlation between productivity and a single rhythm parameter. In step 1, we got a correlation coefficient and a P value for each behavior, rhythm setting, and week. In step 2, we calculated how rhythms changed related to productivity for each behavior sensor weekly (2a) and for each rhythm setting weekly (2b). In step 3, we combined the correlation and importance scores from all 3 weeks. Finally, in step 4, we converted the correlation and importance values into 1 final score for each sensor behavior (4a) and each rhythm setting (4b).



Ethical Considerations

All data collection procedures were approved by an American university’s institutional review board (Carnegie Mellon University; STUDY2016_00000421).

Results

Overview

While correlations between rhythm parameters and productivity scores were moderate across all behavioral sensor features and all 3 weeks (Figure 4), we observed more pronounced relationships between parameters related to regularity in rhythm

models, including SE, that is, deviation of the fitted model parameter from the actual values, percent rhythms (PR and integrated percent rhythm [IPR]) or proportion of variation accounted for by the fitted model, and the significance of the fit (P value and integrated P value [IP]). In addition, the aggregated negative correlation (indicated by the red line) in the majority of these parameters across all 3 weeks indicates lower rhythm irregularity in highly productive students. The rhythm parameters for location features appeared to be dominant in both aggregated correlation coefficients and significance scores, followed by activity and sleep features (Figure 5). In the following, we discuss our observations in detail.

Figure 4. The heat map displays correlations between rhythm parameters and productivity by week. (A) Average correlation coefficients (C-RF) by week (Week-C); (B) Average significance score (S-RF) by week (Week-S). AMP: amplitude; C: correlation coefficients; IP: integrated *P* value; IPR: integrated percent rhythm; MESOR: Midline Statistic of Rhythm; P: *P* value; PHI: acrophase; PR: percent rhythm; RF: random forest; S: significance score.

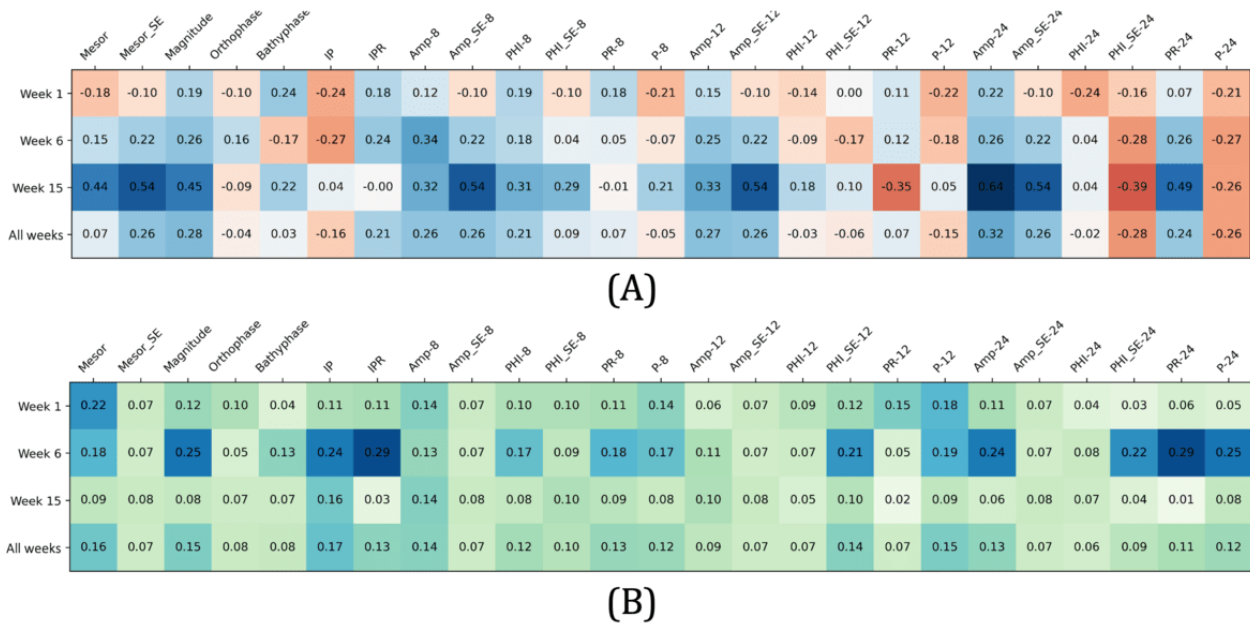
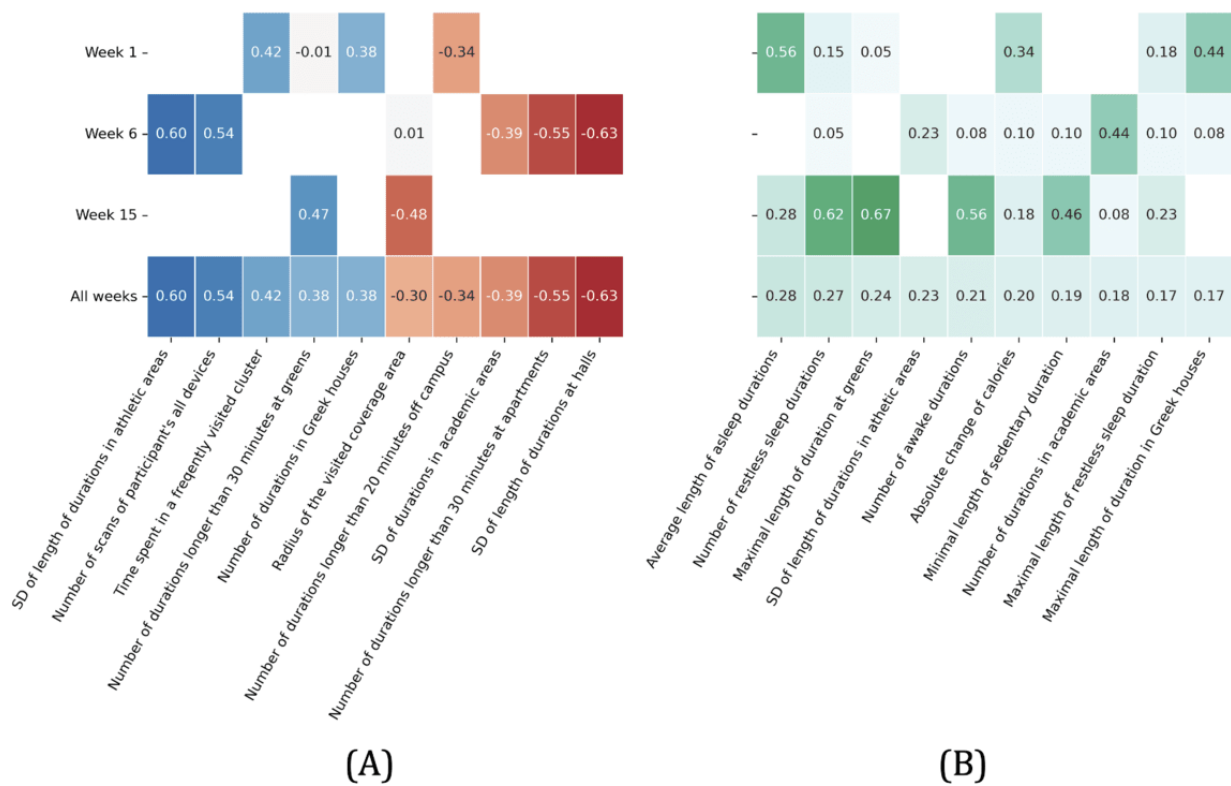


Figure 5. The heat map displays the correlation between sensor features and productivity by week. The left side shows the 10 sensor features with the highest aggregated correlation over all 3 weeks, and the right side shows the 10 sensor features with the highest aggregated significance score over all 3 weeks. The blank cells shown in the figure mean that the relationship is not significant. (A) Average of correlation (C-SF); (B) Significance score of correlation (S-SF). C: correlation coefficients; S: significance score; SF: sensor feature.



Correlation Aggregation of Rhythm Parameters

Overview

The blue and red cells in [Figure 4](#) show the correlation aggregated by week for each rhythm parameter as calculated using equations 8, 9, 11, and 13 in [Multimedia Appendix 2](#). Recall that these formulas aggregate correlation across all sensor features for each rhythm parameter to measure the strength of the correlation between productivity and the rhythm parameter. Blue cells indicate a positive correlation while red cells indicate a negative correlation.

The green cells in [Figure 4](#) show the significance score by week for each rhythm parameter as computed by equations 10 and 12 in [Multimedia Appendix 2](#). These formulas calculate correlation significance across all sensor features for each rhythm parameter to measure the significance of the correlation between productivity and the rhythm parameter. The higher the significance score, the more significant the relationship is.

Week 1

In week 1, the majority of parameters that measure the irregularity of the rhythm models correlate negatively with productivity indicating more stable rhythms in the high productivity group. For example, stronger correlations were observed between productivity and the model fit for the fundamental period (IP; $C=-0.24$), the 24-hour period (P-24; $C=-0.21$), the 12-hour period (P-12; $C=-0.22$), the 8-hour period (P-8; $C=-0.21$), the fundamental PR (IPR; $C = 0.18$), and SE of phase fit for the 24-hour period (PHI_SE-24; $C=-0.16$).

The relationship between regularity in rhythms and productivity is further reinforced by the negative aggregated correlation coefficients for P-24, P-12, P-8, IP, and SE. Specifically, their low values indicate that Cosinor was able to create close fits to the actual data which means more regularity in the actual data corresponds to high productivity. This further demonstrates a lower rhythm variation in highly productive students.

The relationship between lower rhythm variability and higher productivity is also observed in the correlation of MESOR_SE, Amp_SE-8, Amp_SE-12, Amp_SE-24, and PHI_SE-24. The values have a relatively high aggregated significance score compared to other parameters. This means the SE has a more significant relationship with productivity. Given that the SE is also a metric reflecting the irregularity of rhythm models, its negative correlation indicates less irregularity of the rhythm models in high productivity.

The PR parameter also demonstrated a relationship between low rhythm variability and high productivity. A higher PR represents low variability in the actual data. Specifically, the PR of the fundamental, 24-hour, 12-hour, and 8-hour periods all have high positive aggregated correlation coefficients with productivity, indicating lower variability in diurnal activities for the highly productive students.

Week 6

Week 6 (midterm) projected a relatively different pattern. For example, we found positive correlations between productivity and MESOR_SE, Amp_SE-8, Amp_SE-12, and Amp_SE-24.

Since Amp and MESOR are indicative of the intensity and volume of activities, we see that highly productive students performed more intense activity during week 6.

We also found Amp and MESOR have higher SE in the fitted models. This implies higher variability in the intensity of regular patterns during this week. This can be expected due to midterm pressure.

Despite this increased variability of intensity of regular activities, as demonstrated by the positive aggregated correlations of IPR ($C=0.24$) and PR-24 ($C=0.26$) with productivity, we see less irregularity in activity patterns during this week for the highly productive students.

Finally, as in week 1, we see positive correlations between PR and productivity. However, the correlation became more stable in week 6 compared to week 1 with larger aggregated significance scores.

Week 15

Week 15 (the week before finals) showed the strongest correlations. For example, parameters that reflect irregularity in rhythms such as SE (eg, MESOR_SE, Amp_SE, and PHI_SE) show high (mostly positive) correlations with productivity. Parameters characterizing the fitted cyclic model such as MESOR, phase, and Amp also show high (mostly positive) correlations with productivity indicating higher intensity and duration of behavioral activities during this week.

The value of some correlations, however, decreased from weeks 1 and 6 to week 15. For example, the correlation between PRs (eg, IPR, PR_8, and PR_12) and productivity. Given the increased workload activities close to final examinations, the observed irregularity and divergence from the routine patterns are expected.

Despite the decline in the value of some correlations, observations across all 3 weeks still suggest an overall lower irregularity in rhythms among the high-productivity group. For example, there is a consistent negative correlation of the regularity indicators such as P-24, P-12, P-8, PHI-SE-24, PHI-SE-12, PHI-SE-8, and IP. Moreover, parameters representing the phase's characteristics in rhythms including orthophase, bathyphase, PHI-24, PHI-12, and PHI-8 exhibit relatively high aggregated significance scores in all 3 weeks. This means more regularity in phase is more significantly correlated with high productivity. Thus, while further explorations are needed, these observations indicate the importance of rhythm stability in students' productivity.

Correlation Aggregation of Sensor Features

Overview

[Figure 5](#) shows the aggregated correlation and significance scores by week for the top 10 sensor features calculated through equations 2, 3, 4, 5, 6, and 7 in [Multimedia Appendix 1](#). These formulas calculate the aggregated correlation coefficients and significance scores across all rhythm parameters for each sensor feature to measure the strength of the correlation between productivity and behavioral sensor features. Features with higher significance scores have a more significant correlation with

productivity. Overall, location features had a stronger aggregated correlation and significance. The rhythm model for each sensor feature was not consistently associated with productivity in all 3 weeks.

Week 1

In week 1, rhythm parameters for both the time spent in frequently visited places and the frequency of visits in fraternity or sorority houses (places for socializing) showed the highest average positive correlations with productivity. A negative correlation between productivity and off-campus duration was also observed in the rhythm models. Finally, we found patterns of asleep and burned calories to have high significance scores.

Week 6

In week 6, the variance of the length or number of stays in academic areas, halls, and apartments showed high negative aggregated correlations with productivity (the left side of Figure 5), indicating that highly productive students had a stable living and studying environment at home and school. Conversely, the SD of duration in athletic areas was positively correlated with productivity. This indicates higher variability in exercise associated with high productivity. A similar conclusion can be drawn with the data from the aggregated significance score data (the right side of Figure 5).

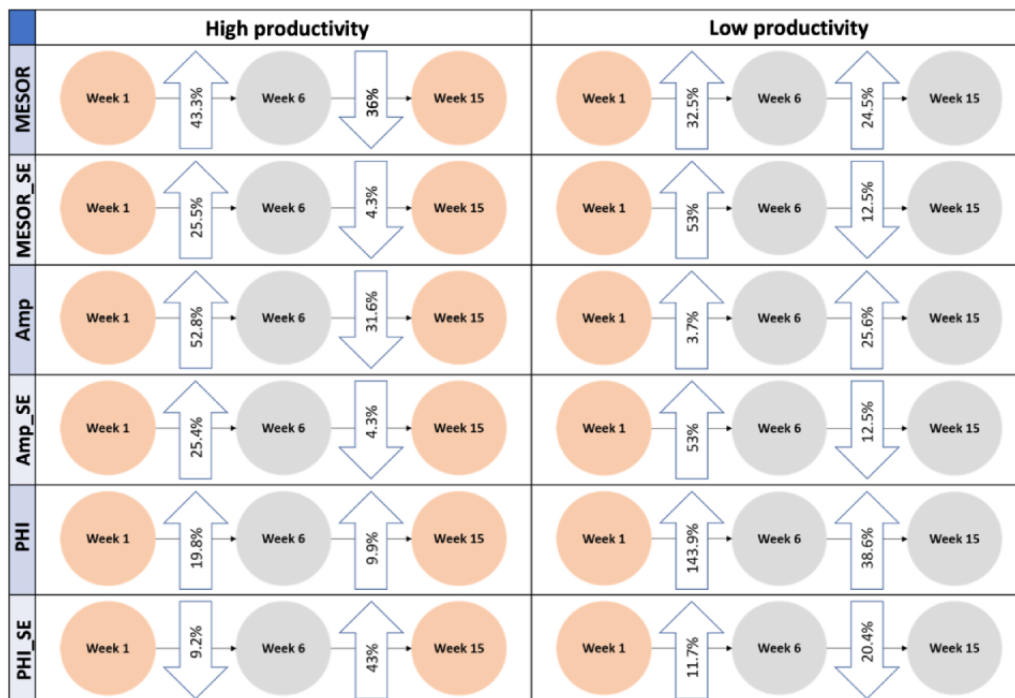
Week 15

In week 15, we observed the highest aggregated significance scores for rhythms of restless sleep duration, awake sleep

duration, time spent at greens, and sedentary duration. On the left side of Figure 5, we see the time spent at greens was positively correlated with productivity, whereas the radius of the visited areas suggests was negatively correlated with productivity. This finding suggests that high-efficiency students reduced their range of activities and spent time outdoors more frequently in week 15.

We further select the “restless sleep” feature to visualize how changes in rhythm parameters reflect the change in productivity for 2 individual students in our sample (Figure 6). The left and right columns in the figure show changes in rhythm parameters between weeks for 1 high- and 1 low-productivity student, respectively. While both students’ productivity levels lowered in week 6, their rhythm parameters of MESOR (SE), Amp (SE), and phase increased from week 1 to 6 with substantially higher variations in the parameters of the low-productive student. After week 6, the student’s productivity in the left column went back to high while MESOR and Amp of their restless sleep rhythm substantially lowered. However, the pattern for the student on the right remained relatively unchanged. As the values of these parameters reflect intensity (Amp and MESOR), duration (phase), and variation (SE), the figure shows that an increase in intensity, duration, and irregularity of restless sleep may be indicative of lower productivity in both students. Although we only look at 2 random participants, the positive and negative changes in rhythm parameters and their accordance with changes in productivity pose an interesting observation and call for further exploration.

Figure 6. Change in restless sleep and productivity patterns of 2 sample students. Orange and gray represent high and low productivity, respectively. The direction of the arrows indicates an increase or decrease of the rhythm parameter values between weeks. AMP: amplitude; MESOR: Midline Estimating Statistic of Rhythm; PHI: acrophase.



Discussion

Principal Findings

In this paper, we analyze cyclic human behavior using passive multimodal mobile sensing data to understand its correlation with work productivity. By creating rhythmic models for each sensor type and employing a multidimensional correlation-based algorithm, we examine the links between biobehavioral rhythms and daily work performance evaluations. Our data are sourced from smartphones and Fitbits of 166 college students, capturing behaviors such as activity, communication, and sleep patterns. The main aim of our analysis is to identify variations in biobehavioral rhythms based on productivity levels and identify specific rhythmic traits associated with them.

To the best of our knowledge, this study pioneers the modeling relationships between daily productivity and biobehavioral rhythms derived from passive sensor data. Notably, we evaluate the capability to model cyclic behavior from detailed phone and Fitbit data. Additionally, we introduce a novel method to measure the correlation and importance of various sensors and rhythms to productivity, which illuminates the connection between rhythmic consistency and different levels of productivity.

Overall, our results showed more rhythm stability in the high-productivity group of students in our sample despite changes in students' workload in different weeks. This observation was especially projected by lower variation accounted for in fitted rhythm models (indicated by PRs and SE parameters) and more significant fit levels (indicated by P parameters) across the weeks. In addition, our correlation analysis of rhythms for each sensor feature showed the significance of consistent patterns in location and sleep to productivity. While encouraging, these results call for more data and analyses to replicate and improve.

Limitations

However, this study was not devoid of limitations. A notable constraint was data quality and its lack of completeness. Inherent issues such as device malfunctions, device misplacement, and time zone travels are usual and expected in mobile and sensor data collection studies. These issues were frequently observed in our data set and contributed to different lengths of time series data for each sensor feature in the modeling step. To address this, we employed data imputation and elimination strategies. The longitudinal repeated-measures design of our study helps mitigate the influence of transient noise or anomalies in the data. By modeling everyone's rhythms across multiple weeks, we reduced the influence of random confounding events. However, we acknowledge that the persistent confounds affecting multiple weeks of data for a given participant could bias their overall rhythms models. We plan to further evaluate our methods on other similar data sets of human behavior such as Tesseract [44], TILES [45], and RAAMPS [46]. We also plan

to extend our study to other groups such as construction workers and office staff in the future.

Few other limitations were imposed by the data set we used in this paper, notably its inclusion of only 3 weeks of noncontinuous self-reported productivity covering the beginning, middle, and end of a semester despite continuous sensor data. Although this was deliberately designed to reduce the burden of frequent self-reports, it limited our ability to model the relationship between productivity and rhythms continuously and throughout the semester. In this study, we incorporated the subjective assessments of daily productivity provided by students through evening surveys. Such survey-based methods are widely recognized in academic research as a standard approach to measure productivity, as evidenced by studies such as Tesseract [44], TILES [45], and RAAMPS [46]. It is worth noting that while subjective measurements might introduce biases, our data indicated that students maintained consistency in their responses over several weeks. Furthermore, by creating individual models for each student's rhythms, we successfully accounted for week-to-week variations, allowing us to assess the relationship between these rhythms and the reported productivity, even considering potential biases. Overall, we were able to test our methods on this data. However, a larger and more longitudinal data set is needed to fully characterize biobehavioral rhythms from mobile data streams and model their relationship with different outcomes.

Given the observational nature of collecting sensor data unobtrusively "in the wild," it is impossible to account for all variables that may impact the data. However, we have taken steps to qualify the potential limitations and strengthen the validity of our digital phenotyping approach within reason. We also suggest further research incorporating both subjective self-reports and sensor data to better characterize confounding contexts. With these caveats articulated, we believe our study maintains substantial value in demonstrating the promise of modeling multidimensional digital phenotypes through passively collected mobile sensor data to advance biobehavioral research.

Conclusion

We explored the feasibility of modeling biobehavioral rhythms from longitudinal multimodal mobile data streams, focusing on college students to identify the relationship between these rhythms and productivity levels. We introduced a multidimensional correlation method to analyze connections between variations in biobehavioral rhythms and productivity. This approach enabled us to observe differences in the longitudinal behavior of high and low-productive students and highlighted that highly productive students encompass more rhythm stability throughout the semester despite variations in workload during different periods. We plan to further evaluate by testing the applicability and adaptability of our methods with diverse data sets. This research paves the way for novel cyber-human systems that align with human beings' biobehavioral rhythms to improve health, well-being, and work performance.

Acknowledgments

This research was supported by the National Science Foundation (NSF) under grant number IIS-1816687.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Correlation between productivity and each sensor feature.

[\[DOCX File, 24 KB - ai_v3i1e47194_app1.docx\]](#)

Multimedia Appendix 2

Correlation between productivity and each rhythm parameter.

[\[DOCX File, 23 KB - ai_v3i1e47194_app2.docx\]](#)

References

1. Doryab A, Dey AK, Kao G, Low C. Modeling biobehavioral rhythms with passive sensing in the wild: a case study to predict readmission risk after pancreatic surgery. *Proc ACM Interact Mob Wearable Ubiquitous Technol* 2019;3(1):1-21. [doi: [10.1145/3314395](#)]
2. Babkoff H, Mikulincer M, Caspy T, Carasso RL, Sing H. The implications of sleep loss for circadian performance accuracy. *Work Stress* 1989;3(1):3-14. [doi: [10.1080/02678378908256875](#)]
3. Folkard S. Circadian performance rhythms: some practical and theoretical implications. *Philos Trans R Soc Lond B Biol Sci* 1990;327(1241):543-553 [FREE Full text] [doi: [10.1098/rstb.1990.0097](#)] [Medline: [1970900](#)]
4. Pope NG. How the time of day affects productivity: evidence from school schedules. *Rev Econ Stat* 2016;98(1):1-11. [doi: [10.1162/rest_a_00525](#)]
5. Smith MR, Eastman CI. Shift work: health, performance and safety problems, traditional countermeasures, and innovative management strategies to reduce circadian misalignment. *Nat Sci Sleep* 2012;4:111-132 [FREE Full text] [doi: [10.2147/NSS.S10372](#)] [Medline: [23620685](#)]
6. Vidacek S, Kaliterna L, Radosević-Vidacek B, Folkard S. Productivity on a weekly rotating shift system: circadian adjustment and sleep deprivation effects? *Ergonomics* 1986;29(12):1583-1590. [doi: [10.1080/00140138608967271](#)] [Medline: [3816750](#)]
7. Yan R, Liu X, Dutcher J, Tumminia M, Villalba D, Cohen S, et al. A computational framework for modeling biobehavioral rhythms from mobile and wearable data streams. *ACM Trans Intell Syst Technol* 2022;13(3):1-27 [FREE Full text] [doi: [10.1145/3510029](#)]
8. Yan R, Ringwald WR, Hernandez JV, Kehl M, Bae SW, Dey AK, et al. Exploratory machine learning modeling of adaptive and maladaptive personality traits from passively sensed behavior. *Future Gener Comput Syst* 2022;132:266-281 [FREE Full text] [doi: [10.1016/j.future.2022.02.010](#)] [Medline: [35342213](#)]
9. Yan R, Doryab A. Towards a computational framework for automated discovery and modeling of biological rhythms from wearable data streams. In: Arai K, editor. *Intelligent Systems and Applications: Proceedings of the 2021 Intelligent Systems Conference (IntelliSys) Volume 3*. Cham: Springer International Publishing; 2021:643-661.
10. Antoniadis EA, Ko CH, Ralph MR, McDonald RJ. Circadian rhythms, aging and memory. *Behav Brain Res* 2000;111(1-2):25-37. [doi: [10.1016/s0166-4328\(00\)00145-5](#)] [Medline: [10840129](#)]
11. Gale JE, Cox HI, Qian J, Block GD, Colwell CS, Matveyenko AV. Disruption of circadian rhythms accelerates development of diabetes through pancreatic beta-cell loss and dysfunction. *J Biol Rhythms* 2011;26(5):423-433 [FREE Full text] [doi: [10.1177/0748730411416341](#)] [Medline: [21921296](#)]
12. Logan RW, McClung CA. Rhythms of life: circadian disruption and brain disorders across the lifespan. *Nat Rev Neurosci* 2019;20(1):49-65 [FREE Full text] [doi: [10.1038/s41583-018-0088-y](#)] [Medline: [30459365](#)]
13. Bellivier F, Geoffroy PA, Etain B, Scott J. Sleep- and circadian rhythm-associated pathways as therapeutic targets in bipolar disorder. *Expert Opin Ther Targets* 2015;19(6):747-763. [doi: [10.1517/14728222.2015.1018822](#)] [Medline: [25726988](#)]
14. Germain A, Kupfer DJ. Circadian rhythm disturbances in depression. *Hum Psychopharmacol* 2008;23(7):571-585 [FREE Full text] [doi: [10.1002/hup.964](#)] [Medline: [18680211](#)]
15. Wulff K, Dijk DJ, Middleton B, Foster RG, Joyce EM. Sleep and circadian rhythm disruption in schizophrenia. *Br J Psychiatry* 2012;200(4):308-316 [FREE Full text] [doi: [10.1192/bjp.bp.111.096321](#)] [Medline: [22194182](#)]
16. Sack RL, Auckley D, Auger RR, Carskadon MA, Wright KP, Vitiello MV, et al. Circadian rhythm sleep disorders: part I, basic principles, shift work and jet lag disorders. *Sleep* 2007;30(11):1460-1483 [FREE Full text] [doi: [10.1093/sleep/30.11.1460](#)] [Medline: [18041480](#)]
17. Valdez P. Circadian rhythms in attention. *Yale J Biol Med* 2019;92(1):81-92 [FREE Full text] [Medline: [30923475](#)]
18. Abdullah S, Matthews M, Murnane EL, Gay G, Choudhury T. Towards circadian computing: "early to bed and early to rise" makes some of us unhealthy and sleep deprived. 2014 Presented at: UbiComp '14: Proceedings of the 2014 ACM

- International Joint Conference on Pervasive and Ubiquitous Computing; September 13-17, 2014; Seattle, Washington p. 673-684. [doi: [10.1145/2632048.2632100](https://doi.org/10.1145/2632048.2632100)]
19. Murmane EL, Abdullah S, Matthews M, Kay M, Kientz JA, Choudhury T, et al. Mobile manifestations of alertness: connecting biological rhythms with patterns of smartphone app use. 2016 Presented at: MobileHCI '16: Proceedings of the 18th International Conference on Human-Computer Interaction with Mobile Devices and Services; September 6-9, 2016; Florence, Italy p. 465-477. [doi: [10.1145/2935334.2935383](https://doi.org/10.1145/2935334.2935383)]
 20. Gloria M, Shamsi TI, Mary C, Paul J, Akane S, Yuliya L. Email duration, batching and self-interruption: patterns of email use on productivity and stress. 2016 Presented at: CHI '16: Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems; May 7-12, 2016; San Jose, California, USA p. 1717-1728. [doi: [10.1145/2858036.2858262](https://doi.org/10.1145/2858036.2858262)]
 21. Garza JL, Cavallari JM, Eijkelhof BHW, Huysmans MA, Thamsuwan O, Johnson PW, et al. Office workers with high effort-reward imbalance and overcommitment have greater decreases in heart rate variability over a 2-h working period. *Int Arch Occup Environ Health* 2015;88(5):565-575. [doi: [10.1007/s00420-014-0983-0](https://doi.org/10.1007/s00420-014-0983-0)] [Medline: [25249418](https://pubmed.ncbi.nlm.nih.gov/25249418/)]
 22. Gatti UC, Migliaccio GC, Bogus SM, Schneider S. Using wearable physiological status monitors for analyzing the physical strain-productivity relationship for construction tasks. *Comput Civ Eng* 2012:577-585. [doi: [10.1061/9780784412343.0073](https://doi.org/10.1061/9780784412343.0073)]
 23. Lee W, Lin KY, Seto E, Migliaccio GC. Wearable sensors for monitoring on-duty and off-duty worker physiological status and activities in construction. *Autom Constr* 2017;83:341-353. [doi: [10.1016/j.autcon.2017.06.012](https://doi.org/10.1016/j.autcon.2017.06.012)]
 24. Punait S, Lewis GF. Theory informed framework for integrating environmental and physiologic data in applications targeting productivity and well-being in workplace. 2019 Presented at: UbiComp/ISWC '19 Adjunct: Adjunct Proceedings of the 2019 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2019 ACM International Symposium on Wearable Computers; September 9-13, 2019; London, United Kingdom p. 179-182. [doi: [10.1145/3341162.3343829](https://doi.org/10.1145/3341162.3343829)]
 25. Aryal A, Ghahramani A, Becerik-Gerber B. Monitoring fatigue in construction workers using physiological measurements. *Autom Constr* 2017;82:154-165. [doi: [10.1016/j.autcon.2017.03.003](https://doi.org/10.1016/j.autcon.2017.03.003)]
 26. Mak CM, Lui YP. The effect of sound on office productivity. *Build Serv Eng Res Technol* 2012;33(3):339-345. [doi: [10.1177/0143624411412253](https://doi.org/10.1177/0143624411412253)]
 27. Seppanen O, Fisk WJ, Faulkner D. Control of temperature for health and productivity in offices. Lawrence Berkeley National Laboratory. 2004. URL: <https://escholarship.org/content/qt39s1m92c/qt39s1m92c.pdf> [accessed 2024-02-23]
 28. Tanabe SI, Nishihara N, Haneda M. Indoor temperature, productivity, and fatigue in office tasks. *HVACR Res* 2007;13(4):623-633. [doi: [10.1080/10789669.2007.10390975](https://doi.org/10.1080/10789669.2007.10390975)]
 29. Wargocki P, Wyon DP, Sundell J, Clausen G, Fanger PO. The effects of outdoor air supply rate in an office on perceived air quality, Sick Building Syndrome (SBS) symptoms and productivity. *Indoor Air* 2000;10(4):222-236 [FREE Full text] [doi: [10.1034/j.1600-0668.2000.010004222.x](https://doi.org/10.1034/j.1600-0668.2000.010004222.x)] [Medline: [11089327](https://pubmed.ncbi.nlm.nih.gov/11089327/)]
 30. Mirjafari S, Masaba K, Grover T, Wang W, Audia P, Campbell AT, et al. Differentiating higher and lower job performers in the workplace using mobile sensing. *Proc ACM Interact Mob Wearable Ubiquitous Technol* 2019;3(2):1-24. [doi: [10.1145/3328908](https://doi.org/10.1145/3328908)]
 31. van Vugt M. Using biometric sensors to measure productivity. In: Sadowski C, Zimmermann T, editors. *Rethinking Productivity in Software Engineering*. Berkeley, CA: Apress; 2019:159-167.
 32. Das Swain V, Saha K, Rajvanshy H, Sirigiri A, Gregg JM, Lin S, et al. A multisensor person-centered approach to understand the role of daily activities in job performance with organizational personas. *Proc ACM Interact Mob Wearable Ubiquitous Technol* 2020;3(4):1-27. [doi: [10.1145/3369828](https://doi.org/10.1145/3369828)]
 33. Ferreira D, Kostakos V, Dey AK. AWARE: mobile context instrumentation framework. *Front ICT* 2015;2:6 [FREE Full text] [doi: [10.3389/fict.2015.00006](https://doi.org/10.3389/fict.2015.00006)]
 34. Doryab A, Chikarsel P, Liu X, Dey AK. Extraction of behavioral features from smartphone and wearable data. ArXiv Preprint posted online on December 18, 2018 [FREE Full text] [doi: [10.48550/arXiv.1812.10394](https://doi.org/10.48550/arXiv.1812.10394)]
 35. Donders ART, van der Heijden GJMG, Stijnen T, Moons KGM. Review: a gentle introduction to imputation of missing values. *J Clin Epidemiol* 2006;59(10):1087-1091 [FREE Full text] [doi: [10.1016/j.jclinepi.2006.01.014](https://doi.org/10.1016/j.jclinepi.2006.01.014)] [Medline: [16980149](https://pubmed.ncbi.nlm.nih.gov/16980149/)]
 36. Lin WC, Tsai CF. Missing value imputation: a review and analysis of the literature (2006–2017). *Artif Intell Rev* 2020;53(2):1487-1509. [doi: [10.1007/s10462-019-09709-4](https://doi.org/10.1007/s10462-019-09709-4)]
 37. Halberg F. Chronobiology. *Annu Rev Physiol* 1969;31:675-726. [doi: [10.1146/annurev.ph.31.030169.003331](https://doi.org/10.1146/annurev.ph.31.030169.003331)] [Medline: [4885778](https://pubmed.ncbi.nlm.nih.gov/4885778/)]
 38. Halberg F, Engeli M, Hamburger C, Hillman D. Spectral resolution of low-frequency, small-amplitude rhythms in excreted 17-ketosteroids; probable androgen-induced circaseptan desynchronization. *Acta Endocrinol (Copenh)* 1965;50(4_Suppl):S5-S54. [doi: [10.1530/acta.0.050s0005](https://doi.org/10.1530/acta.0.050s0005)] [Medline: [5898281](https://pubmed.ncbi.nlm.nih.gov/5898281/)]
 39. Cornelissen G. Cosinor-based rhythmometry. *Theor Biol Med Model* 2014;11(1):16 [FREE Full text] [doi: [10.1186/1742-4682-11-16](https://doi.org/10.1186/1742-4682-11-16)] [Medline: [24725531](https://pubmed.ncbi.nlm.nih.gov/24725531/)]
 40. Fernández JR, Hermida RC, Mojón A. Chronobiological analysis techniques. Application to blood pressure. *Philos Trans A Math Phys Eng Sci* 2009;367(1887):431-445 [FREE Full text] [doi: [10.1098/rsta.2008.0231](https://doi.org/10.1098/rsta.2008.0231)] [Medline: [18940774](https://pubmed.ncbi.nlm.nih.gov/18940774/)]
 41. Gierke CL, Cornelissen G. Chronomics Analysis Toolkit (CATkit). *Biol Rhythm Res* 2016;47(2):163-181. [doi: [10.1080/09291016.2015.1094965](https://doi.org/10.1080/09291016.2015.1094965)]

42. Borenstein M, Hedges LV, Higgins JPT, Rothstein HR. Introduction to Meta-Analysis, 2nd Edition. Hoboken, NJ: John Wiley & Sons; 2021.
43. Fisher RA. Statistical methods for research workers. In: Kotz S, Johnson NL, editors. Breakthroughs in Statistics. Springer Series in Statistics. New York, NY: Springer; 1992:66-70.
44. Mattingly SM, Gregg JM, Audia P, Bayraktaroglu AE, Campbell AT, Chawla NV, et al. The Tesseract project: large-scale, longitudinal, in situ, multimodal sensing of information workers. 2019 Presented at: CHI EA '19: Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems; May 4-9, 2019; Glasgow, Scotland, UK p. 1-8. [doi: [10.1145/3290607.3299041](https://doi.org/10.1145/3290607.3299041)]
45. Mundnich K, Booth BM, L'Hommedieu M, Feng T, Girault B, L'Hommedieu J, et al. TILES-2018, a longitudinal physiologic and behavioral data set of hospital workers. Sci Data 2020;7(1):354. [doi: [10.1038/s41597-020-00655-3](https://doi.org/10.1038/s41597-020-00655-3)] [Medline: [33067468](https://pubmed.ncbi.nlm.nih.gov/33067468/)]
46. Danvers A, Notaro G, Kraft A, Baraniecki L, Baranski E, Alexander W, et al. Rapid Automatic and Adaptive Models for Performance Prediction (RAAMP2) Dataset. Center for Open Science. 2020. URL: <https://osf.io/9e86j/> [accessed 2024-02-23]

Abbreviations

Amp: amplitude

IP: integrated *P* value

IPR: integrated percent rhythm

MESOR: Midline Estimating Statistic of Rhythm

PHI: acrophase

PR: percent rhythm

Edited by J Sun; submitted 12.03.23; peer-reviewed by J Wang, PB Chandrashekar, E Sükei; comments to author 31.07.23; revised version received 31.10.23; accepted 15.02.24; published 18.04.24.

Please cite as:

Yan R, Liu X, Dutcher JM, Tumminia MJ, Villalba D, Cohen S, Creswell JD, Creswell K, Mankoff J, Dey AK, Doryab A
Identifying Links Between Productivity and Biobehavioral Rhythms Modeled From Multimodal Sensor Streams: Exploratory Quantitative Study

JMIR AI 2024;3:e47194

URL: <https://ai.jmir.org/2024/1/e47194>

doi: [10.2196/47194](https://doi.org/10.2196/47194)

PMID:

©Runze Yan, Xinwen Liu, Janine M Dutcher, Michael J Tumminia, Daniella Villalba, Sheldon Cohen, John D Creswell, Kasey Creswell, Jennifer Mankoff, Anind K Dey, Afsaneh Doryab. Originally published in JMIR AI (<https://ai.jmir.org>), 18.04.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR AI, is properly cited. The complete bibliographic information, a link to the original publication on <https://www.ai.jmir.org/>, as well as this copyright and license information must be included.

Original Paper

Impact of a Digital Scribe System on Clinical Documentation Time and Quality: Usability Study

Marieke Meija van Buchem¹, BS, MS; Ilse M J Kant², BS, MS, PhD; Liza King³, BA, MS; Jacqueline Kazmaier³, BEng, MS, MEng, PhD; Ewout W Steyerberg⁴, BS, MS, PhD; Martijn P Bauer⁵, BS, MS, PhD

¹CAIRELab (Clinical AI Implementation and Research Lab), Leiden University Medical Center, Leiden, Netherlands

²Department of Digital Health, University Medical Center Utrecht, Utrecht, Netherlands

³Autoscriber B.V., Eindhoven, Netherlands

⁴Department of Biomedical Data Sciences, Leiden University Medical Center, Leiden, Netherlands

⁵Department of Internal Medicine, Leiden University Medical Center, Leiden, Netherlands

Corresponding Author:

Marieke Meija van Buchem, BS, MS

CAIRELab (Clinical AI Implementation and Research Lab)

Leiden University Medical Center

Albinusdreef 2

Leiden, 2333 ZN

Netherlands

Phone: 31 615609183

Email: m.m.van_buchem@lumc.nl

Abstract

Background: Physicians spend approximately half of their time on administrative tasks, which is one of the leading causes of physician burnout and decreased work satisfaction. The implementation of natural language processing–assisted clinical documentation tools may provide a solution.

Objective: This study investigates the impact of a commercially available Dutch digital scribe system on clinical documentation efficiency and quality.

Methods: Medical students with experience in clinical practice and documentation (n=22) created a total of 430 summaries of mock consultations and recorded the time they spent on this task. The consultations were summarized using 3 methods: manual summaries, fully automated summaries, and automated summaries with manual editing. We then randomly reassigned the summaries and evaluated their quality using a modified version of the Physician Documentation Quality Instrument (PDQI-9). We compared the differences between the 3 methods in descriptive statistics, quantitative text metrics (word count and lexical diversity), the PDQI-9, Recall-Oriented Understudy for Gisting Evaluation scores, and BERTScore.

Results: The median time for manual summarization was 202 seconds against 186 seconds for editing an automatic summary. Without editing, the automatic summaries attained a poorer PDQI-9 score than manual summaries (median PDQI-9 score 25 vs 31, $P < .001$, ANOVA test). Automatic summaries were found to have higher word counts but lower lexical diversity than manual summaries ($P < .001$, independent t test). The study revealed variable impacts on PDQI-9 scores and summarization time across individuals. Generally, students viewed the digital scribe system as a potentially useful tool, noting its ease of use and time-saving potential, though some criticized the summaries for their greater length and rigid structure.

Conclusions: This study highlights the potential of digital scribes in improving clinical documentation processes by offering a first summary draft for physicians to edit, thereby reducing documentation time without compromising the quality of patient records. Furthermore, digital scribes may be more beneficial to some physicians than to others and could play a role in improving the reusability of clinical documentation. Future studies should focus on the impact and quality of such a system when used by physicians in clinical practice.

(JMIR AI 2024;3:e60020) doi:[10.2196/60020](https://doi.org/10.2196/60020)

KEYWORDS

large language model; large language models; LLM; LLMs; natural language processing; NLP; deep learning; pilot study; pilot studies; implementation; machine learning; ML; artificial intelligence; AI; algorithm; algorithms; model; models; analytics;

practical model; practical models; automation; automate; documentation; documentation time; documentation quality; clinical documentation

Introduction

In recent years, the issue of burnout among physicians has been increasingly recognized within the health care sector. A survey conducted in 2017 involving 5000 physicians in the United States found that 44% exhibited at least 1 sign of burnout [1]. In response to this issue, the National Academy of Medicine established a committee dedicated to enhancing patient care through the promotion of physician well-being. The committee produced a detailed report titled Taking Action Against Clinician Burnout, which outlines the causes of burnout among physicians. A significant cause identified is the growing administrative workload [2]. The introduction of the electronic health record (EHR) has led to physicians spending up to half of their working hours on administrative duties [3-5]. Such tasks have been shown to lower job satisfaction for physicians [6] and negatively impact the physician-patient relationship [7]. Additionally, research linking the use of EHR to burnout indicates that physicians spending more time on EHR, particularly outside of regular hours, face a greater risk of experiencing burnout [8,9].

Recent advances in natural language processing (NLP) have created the possibility of automating some of these administrative tasks. One of these promises is the creation of the so-called “digital scribe.” Such a system, first described in 2018, automatically records, transcribes, and summarizes the clinical encounter [10,11]. A scoping review from 2022 presented an overview of the capabilities of digital scribes at that point in time, and showed that none of these systems had the full capability of a digital scribe [12]. The introduction of large language models has disrupted this field, with many papers describing their potential value in clinical note generation and multiple companies now offering digital scribe systems [13-15]. However, an evaluation on the potential impact of such a system on documentation time, including the assessment of quality and user experiences is not available to date. A thorough, prospective investigation of digital scribe performance and impact on routine practice is necessary to ensure the safety and effectiveness of the system. The aim of the current study is to assess the potential impact on the time spent and quality of medical summaries using a Dutch, commercially available digital scribe system.

Methods

Data

Our data set consisted of 27 recordings of mock consultations between physicians and nonmedical individuals. The consultations were structured around 26 vignettes, created by an internist. These vignettes delineated a set of symptoms, with a focus on various presentations of chest pain. Nonmedical individuals, assuming the role of patients, were provided with these vignettes. They were encouraged to develop and present a narrative surrounding the described symptoms. The participating physicians, all specialists in internal medicine from the Leiden University Medical Center, engaged with these simulated patients, applying their expertise to the scenarios presented. The average duration of the consultations was 293 (IQR 189-398) seconds.

Participants

In total, 21 medical students with experience in clinical practice and clinical documentation from Leiden University Medical Center consented to participate in our study. All students had a bachelor’s degree in medicine and completed a course in clinical documentation. The students received a compensation of €100 (US \$111) for their participation.

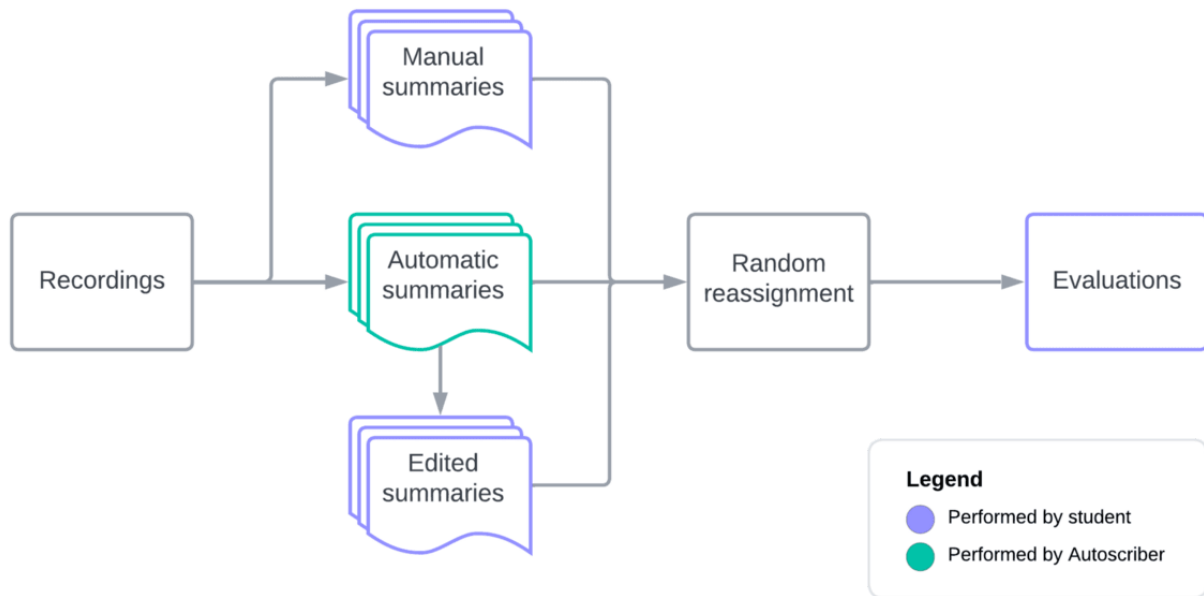
Autoscriber

Autoscriber is a web-based software application that transcribes and summarizes medical conversations (currently with support for Dutch, English, and German). The pipeline uses a transformer-based speech-to-text model, fine-tuned on proprietary clinical data for transcription and a mixture of large language models such as GPT-3.5 and GPT-4, combined with a tailored prompt structure and additional rules for summarization. The tool also has self-learning functionality, which was not evaluated in this study for practical reasons.

Summarization

All students summarized 4 consultations manually, then 8 consultations using Autoscriber, and finally 4 consultations manually to minimize a learning effect (see Figure 1). In total, students summarized 16 unique consultations.

Figure 1. Flowchart showing the 3 different summarization methods and consecutive evaluation.



Manual Summarization

Students were asked to listen to the full recording, making some notes using pen and paper. At the end of the recording, they started timing and summarized the consultation on the computer. When finished, they recorded the total time spent summarizing.

Automatic Summarization

For the 8 consultations summarized using Autoscriber, the setup was similar. However, students first opened the Autoscriber application and, while listening to the recording, also recorded the consultation with Autoscriber. Once Autoscriber had created an automatic summary, students started timing and edited the automatic summary. Finally, they uploaded both the automatic summary and the edited summary, including the total time they spent editing.

Evaluation

Once all summaries were created, the manual, automatic, and edited summaries were randomly reassigned to other students, who were blinded for the method used to create the summary. Students first listened to the full recording, and then evaluated the related summaries using a modified version of the Physician Documentation Quality Instrument (PDQI-9) [16]. The PDQI-9 is a validated evaluation instrument for assessing the quality of clinical documentation, consisting of 9 questions. We removed question 1 (up-to-date: the note contains the most recent test results and recommendations) and 8 (synthesized: the note reflects the author's understanding of the patient's status and ability to develop a plan of care) for our study, as these could not be answered in the current setup. We translated the questions into Dutch, which were reviewed by one clinician (MB). Per recording, we selected the manual summary with the highest PDQI-9 score as the reference standard summary.

At the end of the study, we asked students about their experience with Autoscriber, what was positive, what should be improved,

and if they would want to use Autoscriber in their work. For a more in-depth view of the differences between the automatic and edited summaries, we prompted ChatGPT (paid version, GPT-4) to assess the differences. The prompt was created iteratively using PromptPerfect until the format of the answer was satisfactory. We then ran the prompt several times to check for internal consistency. Two researchers (MB and MvB) manually checked the answers provided by ChatGPT.

Data Analysis

Preprocessing

For every summary, we calculated the total word count and the lexical diversity. Furthermore, to compare the automatic summaries to their edited counterparts we calculated the number of insertions, deletions, the Recall-Oriented Understudy for Gisting Evaluation (ROUGE)-1 and ROUGE-L score [17], and the BERTScore metric [18]. The ROUGE-1 score calculates the overlap in words between 2 texts. The ROUGE-L score calculates the longest common subsequence. The BERTScore metric uses contextual embeddings to compare words between 2 texts.

Power Analysis

To ensure the study was adequately powered to detect a large effect size (Cohen $f=0.4$) between 3 groups with an alpha level of 0.05 and a power of 95%, a power analysis was conducted using the `FTestAnovaPower` function from the `statsmodels` library in Python. This analysis assumed equal group sizes and did not account for potential correlations among repeated measures.

Statistical Analysis

The differences between the automatic and associated edited summaries were tested using a paired t test. To compare the differences in summaries per recording, we selected the manual summary with the highest PDQI-9 score as the reference

standard. We then calculated the ROUGE-1 and ROUGE-L scores for all the other manual, automatic, and edited summaries. The differences in word count, lexical diversity, PDQI-9 score, and ROUGE scores between the 3 methods was tested using one-way ANOVA and, if the *P*-value was below .05, followed by Tukey Honestly Significant Difference test. To assess the possibility of a learning effect, we compared the first and second batch of manual summaries on time spent creating the summary and PDQI-9 score using a paired sample *t* test. We used Python for the analysis, using the “statsmodels” and the “scipy” package.

Ethical Considerations

This study was conducted in accordance with the Declaration of Helsinki. For the purposes of this study, ethics approval was not applicable as the research did not include actual patients or any personal or sensitive information. All students involved in the study were informed about the purpose of the research, the use of the data, and gave their informed consent to participate in the study under these conditions.

Results

The power analysis indicated that a sample size of approximately 100 participants per group would be necessary to achieve the desired power of 95% for detecting a large effect size among the 3 groups under the specified conditions. In total, we collected 156 manual summaries, 137 automatic summaries, and 137 edited summaries from 21 students. A difference in the total number of manual, automatic, and edited summaries occurred because 3 students dropped out of the study due to time restraints. Table 1 shows an example of a manual, automatic, and edited summary of the same recording. 18 students completed the evaluation phase of the study. The median time students spent creating or editing the summaries was 186 seconds (IQR 109-267). Summaries had a median length of 129 (IQR 91-172) words. On average, summaries had a median PDQI-9 score of 28.5 (IQR 25-32) out of a maximum of 35 (Table 2). Multimedia Appendix 1 shows an extended version of Table 2, including the results of the Tukey Honestly Significant Difference test. There was a difference in time spent on manually summarizing the first batch and the second batch, with a median of 246 (IQR 137-311) and 188 (IQR 118-226), respectively, (*P*=.004). However, there was no difference in PDQI-9 score between these 2 batches.

Table 1. An example of a manual, automatic, and edited summary of the same recording.

Manual summary (translated)	Automatic summary	Edited summary
<ul style="list-style-type: none"> For a few days now, tearing pain in the chest with radiation to behind the shoulder blades. Was sitting on the couch watching TV, pain started acutely, felt like a tear and is continuously present. No complaints of tingling, numbness, or cold hands. No loss of function in arm or leg. No palpitations. Not nauseous. Has not had similar complaints before. The patient is on the “edge” of hypercholesterolemia, does not use medication. Measures blood pressure regularly for donation, pressure regular, last 120/75. The patient has never smoked. Drinks alcohol on the weekend. Family: Father had congenital vascular problems, began having heart problems at the beginning of his 50s, also had heart valve problems, died of heart failure. Brother had a heart attack at age 46, was stented. Both brothers have hypercholesterolemia. No connective tissue disorders in the family. The patient is worried due to familial CVD (Cardiovascular Disease). 	<ul style="list-style-type: none"> Reason for referral: <ul style="list-style-type: none"> Main complaint: Persistent chest pain, feels as if something is tearing and radiates to behind the shoulder blades. History: <ul style="list-style-type: none"> Chest pain started a few days ago while calmly watching TV. The pain remains constantly present since the beginning. The pain feels as if something is tearing and radiates to the back of the shoulder blades. There are heart problems in the family (patient’s father and brother). The patient’s father had a poor vascular system and a leaking heart valve, and died of a heart failure. The patient’s brother had a heart attack at the age of 46 and was stented. Social history: The patient does not smoke and drinks moderately alcohol on the weekend. Current medication: The patient does not use any medication. Physical examination: Additional investigation: <ul style="list-style-type: none"> Blood pressure is regularly measured and is about 75 over 120. The patient gives blood every six months. Diagnosis: Treatment plan: Requested examinations: Informed consent: 	<ul style="list-style-type: none"> For several days now, tearing pain in the chest, radiating to between the shoulder blades. Chest pain started a few days ago while calmly watching TV and has been stable since then. No palpitations, not nauseous, no tingling or numb feeling. Blood pressure is regularly measured and is about 75 over 120. Cholesterol is good. Patient is worried because of family history. Fam: father had heart problems, brother had a heart attack at a young age, hypercholesterolemia, no connective tissue diseases. Intox: no smoking, alcohol on weekends in moderation Med: none

Table 2. Descriptive statistics of the different methods and associated *P* values.

Metrics	Manual (n=156), median (IQR)	AS edited (n=137), median (IQR)	AS (n=137), median (IQR)	<i>P</i> value (ANOVA)
Time spent on summary (seconds)	202 (128-286)	152 (93-244)	0 (0-0)	<.001
Word count	101 (67-141)	137 (96-194)	148 (116-180)	<.001
Lexical diversity	0.68 (0.63-0.74)	0.61 (0.56-0.66)	0.59 (0.53-0.63)	<.001
PDQI-9^a score				
Overall	31 (27-33)	29 (26-33)	25 (22-28)	<.001
Accurate	5 (4-5)	5 (4-5)	4 (2-5)	<.001
Thorough	4 (4-5)	4 (4-5)	3 (2-4)	<.001
Useful	5 (4-5)	4 (4-5)	4 (3-4)	<.001
Organized	4 (3-5)	4 (3-5)	4 (3-4)	.01
Comprehensible	5 (4-5)	5 (4-5)	4 (3-5)	<.001
Succinct	5 (4-5)	4 (2-5)	3 (2-4)	<.001
Internally consistent	5 (4-5)	5 (4-5)	5 (4-5)	<.001
ROUGE ^{b,c} -1 <i>F</i> ₁ -score	47.3 (42.5-56.4)	40.6 (35.0-45.4)	32.3 (27.0-37.4)	<.001
ROUGE-L <i>F</i> ₁ -score	29.4 (23.7-37.6)	23.4 (20.6-27.5)	19.6 (15.7-23.5)	<.001
BERTScore ^c <i>F</i> ₁ -score	74.6 (71.9-77.0)	71.6 (69.5-73.7)	68.6 (67.5-70.3)	<.001

^aPDQI-9: Physician Documentation Quality Instrument.

^bROUGE: Recall-Oriented Understudy for Gisting Evaluation.

^cTo calculate the ROUGE score and BERTScore, the highest scoring manual summary was taken as the reference standard. These summaries were taken out of the data set when calculating the average ROUGE scores.

Comparison Between Automatic and Corresponding Edited Summaries

Students inserted a median of 45 (IQR 27-82) words and deleted 46 (IQR 27-80) words. The edits led to a median increase in PDQI-9 score of 4.0 (IQR 1-8). The median ROUGE-1 *F*₁ score between the automatic and their corresponding edited summaries was 73.3 (IQR 61.0-84.4), the ROUGE-L *F*₁ score was 67.4 (IQR 50.0-80.5), and the BERTScore *F*₁ was 84.1 (IQR 79.0-89.4).

ChatGPT assessed the differences between automatic summaries and their edited counterparts on the following aspects: language use and precision, clarity and detail, coherence and flow, structural differences, stylistic variations, and the most common deletions and insertions. The final prompt can be seen in [Multimedia Appendix 2](#). See [Table 3](#) for the observations per aspect. The assessment by ChatGPT aligned with the sample analysis performed by the researchers. Furthermore, similar aspects were mentioned by the students.

Table 3. Differences between automatic and edited summaries, as assessed by ChatGPT.

Aspect	Automatic summaries	Edited summaries	Observations
Language use and precision	Generally simplistic and formulaic language. For example, "Chest pain started a few days ago while quietly watching TV."	More sophisticated and precise language. Example: "Since a few days tearing chest pain radiating to between the shoulder blades."	Human editors refine the language to be more precise and contextually appropriate.
Clarity and detail	Often vague, lacking specific details. For instance, "Patient has had persistent watery diarrhea since one week."	Provide clearer, more detailed descriptions. Example: "Patient has had persistent watery diarrhea for a week with a frequency of ten times a day."	Human editing enhances clarity by adding relevant details that were omitted in the automatic summaries.
Coherence and flow	Sometimes disjointed or lacking in logical flow. Example: "The chest pain started suddenly and has been continuously present since it started."	Better structured, with a smoother flow of ideas. Example: "The patient complains of sudden and persistent chest pain that started several days ago."	Human editors improve the coherence, making the summaries easier to follow.
Structural differences	Tend to follow a predictable structure, possibly template-based.	More varied structures, adapted to the content's needs.	Human editing allows for more flexible structuring, tailored to the specific summary.
Stylistic variations	Limited stylistic variations, often repetitive.	Display a wider range of styles, adapting to the tone and context.	Human editors introduce stylistic diversity, making each summary more unique.
Most common deletions			Redundant phrases, overly general statements.
Most common insertions			Specific details, clarifying phrases, and contextual information.

Differences Per Student

Using Autoscriber had a different effect per student. For 8 out of 18 students, using Autoscriber was associated with a decrease in PDQI-9 score, while for the other students the difference in PDQI-9 score between manual and automatic summaries had a *P* value above .05. For 5 students, editing the automatic summary took more time than manually creating a summary, although these differences were not significant. For 3 students, editing the automatic summary led to a decrease in time spent on summarizing, with a *P* value lower than .05. See [Multimedia Appendix 3](#) for the full overview.

Experiences With the Use of Autoscriber

Students were generally very positive about using Autoscriber, mentioning that it was nice or interesting to use (n=9), easy and

simple in use (n=6), and that they believed in the potential of such a tool (n=4). Four students mentioned the automatic summary exceeded their expectations, while 4 other students said the quality of the summary was insufficient due to errors and the amount of time needed to make edits. A specific error that was mentioned multiple times was that the summary did not include negative symptoms (eg, the absence of shortness of breath). Three students mentioned the tool did not always work: it would sometimes load for a very long time or get stuck while generating the summary. This was due to limitations in graphics processing unit capacity at that time. See [Table 4](#) for the positive aspects and points of improvement mentioned by the students. A majority of students (12/18, 67%) would want to use the application during their work. The other students (6/18, 33%) said they would want to use the application if improvements were made.

Table 4. Themes most often described by students about the positive aspects and points of improvement.

Mentioned aspects	Count
Positive	
Easy to use	5
Good accuracy, eg, amount of details, good use of language, low amount of errors, inclusion of important symptoms	5
Summary fairly complete	4
Saves time	4
Well-structured view	4
Nice to have something to start with, without typing	3
Negative	
Structure does not align with preferences, eg, headings unclear, illogical structure, does not align with style	6
Wordy/lengthy	5
Relevant information missing, eg, details, absence of symptoms	5
Comments on language use, eg, use of nonstandard words, vague descriptions, too literal, absence of common abbreviations	5
Duration of summarization time	3
Presence of irrelevant information	2

Discussion

Principal Findings

In this impact study, we extensively evaluated the efficacy of Autoscriber, a Dutch digital scribe system, in enhancing the clinical documentation process in a pilot setting. A group of trained medical students summarized clinical conversations with and without the tool. We found differences between automatic and manual summaries in time spent on the summary, the word count, lexical diversity, and qualitative aspects such as accurateness and usefulness. These differences decreased after students edited the automatic summaries. During editing, medical students most often added context and details, while removing overly general statements and irrelevant text. Most were positive about using the tool, although some mentioned the summaries were lengthy and the structure did not always align with their preferences.

As the first impact study of a fully functioning digital scribe system, we provide some interesting insights into the possible future of digital scribes in health care. First of all, we show that a collaboration between the system and the students leads to the best results at this point in time, with a decrease in time spent on summarizing in combination with a similar quality when compared to manual summarization. We believe the current setting might even provide an overestimation of the quality of the manual summaries: the students did not have a time cap for creating the summaries, while in clinical practice, physicians often have to create a summary during or in between consultations. Furthermore, multiple studies show a negative association between seniority of a physician and the completeness of a medical record [19-21]. Taking this into account, we see the potential in using a digital scribe system that provides a first draft, which the physician then edits. In the

current setup, this collaboration led to a decrease in time spent summarizing, while keeping the quality of the summary on par.

When looking at the differences between the 3 methods, the higher word count and lower lexical diversity in the automatic summaries compared to the manual summaries stood out. Two previous studies compared human and ChatGPT-written medical texts and reported similar results [22,23]. Furthermore, one of these studies reported human texts contained more specific content, which we found as well. These aspects are essential to improve in future versions, as they directly link to the quality of a summary in terms of succinctness and thoroughness. An increased summary length could lead to an increase in time spent reading or analyzing summaries downstream in the clinical process. However, a small decrease in lexical diversity in combination with a more structured summary could also be seen as a step toward standardization of medical summaries. This aspect is becoming more important since clinical documentation is increasingly reused for other purposes, such as research and quality measurements. Furthermore, previous studies show that structured documentation leads to increased note quality [24], which in turn has been shown to positively affect the quality of care [25-27]. These potential effects have to be studied in future research.

We found large differences in the effect of using Autoscriber on PDQI-9 score and time spent summarizing between students. While using Autoscriber decreased the time spent on finalizing the summary for most students, there were a few students who spent more time on editing the automatic summary than on manually creating a summary. Furthermore, the difference in PDQI-9 score between manual and automatic summaries differed greatly between students. This result is highly relevant, as it shows that the added value of using a digital scribe differs per user. Future studies should investigate which users could

gain most benefit in using a digital scribe, taking into account age, specialty, the ability to type blindly, and other factors that might impact the added value on a personal level.

Strengths and Limitations

This impact study on a digital scribe system for clinical conversations presents a novel exploration into the practical application of such technology. Since the introduction of ChatGPT, many papers have described the potential of using ChatGPT and other large language models in health care. While their potential is clear, these models have still to prove their actual clinical value. This study takes a first step in gaining a better view of the potential effects such a digital scribe system could have on the documentation process, especially in interaction with the user. Apart from quantitative analyses, we also included several different qualitative analyses, providing a more in-depth view of the differences between the summaries and the experiences of the students. These results are highly relevant for researchers and companies developing digital scribes as well as health care organizations considering using a digital scribe in the near future.

One limitation is the setup of our study, which is not fully representative of clinical practice. Specifically, our reliance on medical students listening to prerecorded mock consultations does not fully capture the dynamic and often unpredictable nature of real-time clinical interactions. The controlled environment of our study does not account for the varied technological, environmental, and personal factors that can influence the use and effectiveness of digital scribe systems in live clinical environments. However, this approach allowed us to isolate and evaluate the impact on summarization time and differences in summary between the 3 methods. Future research should aim to incorporate real clinical interactions to validate and extend our findings.

Another limitation is the lack of a reference summary per consultation. To calculate the ROUGE scores, we designated the highest scoring manual summary as the reference standard per consultation. This method suffices for the current pilot study; however, it brings up the bigger issue of summary evaluation metrics. The ROUGE score remains the most used metric, while this metric only measures exact overlap in words and is, thus, very sensitive to the choice of reference summaries [28]. Because of this limitation, we added the BERTScore metric, which has been shown to correlate better with human evaluations [18]. However, the overall lack of a standard for clinical documentation still poses a considerable challenge for the objective assessment of summarization efficacy of digital scribes. This underscores the necessity for developing more

sophisticated evaluation methods, especially with the arrival of large language models in health care.

Future Implications

Our findings underscore the promising potential of integrating digital scribe technologies like Autoscriber within clinical settings to alleviate the administrative burdens faced by health care professionals. Future clinical impact studies are imperative to explore the broader effects of digital scribes on the physician-patient interaction, documentation accuracy, and overall health care delivery efficiency. These studies should aim to evaluate the real-world applicability of digital scribes, including their impact on clinical workflow, quality of care, and patient satisfaction. Especially the latter, which has not received sufficient attention up to now, should be the focus of future research to ensure the physician-patient relationship is not harmed. Additionally, exploring the customization of digital scribe systems to fit the specific needs and preferences of individual physicians or specialties could enhance user adoption and effectiveness. As the field of large language models is developing at a fast rate and digital scribes will improve quickly, repeated or continuous evaluation of these systems is necessary. A recent study described the development and evaluation of a chat-based diagnostic conversational agent [29]. This agent outperformed primary health care providers in both diagnosis and the development of a treatment plan. The introduction of digital scribes in clinical practice could eventually lead to similar support during the clinical encounter, where the digital scribe might suggest additional follow-up questions or provide a differential diagnosis. Ultimately, the goal is to seamlessly integrate digital scribes into clinical practice, ensuring they enhance patient care and physician well-being.

Conclusions

This study explores the impact of a Dutch digital scribe system on the clinical documentation process, offering significant insights into its potential to enhance physicians' experience. By demonstrating the use of the system in reducing summarization time while maintaining summary quality through collaborative editing, our research highlights the potential of digital scribe systems in addressing the challenges of clinical documentation. Despite the limitations related to the representativeness of our pilot setup and the evaluation of summary quality, the positive outcomes suggest a promising avenue for future research and development. Further studies, particularly those involving real-world clinical settings, are essential to fully understand the implications of digital scribes on the physician-patient dynamic and health care delivery.

Acknowledgments

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Data Availability

The majority of the data supporting the findings of this study are included in the manuscript. Additional data sets generated during and analyzed during the current study are not publicly available but are available from the corresponding author on reasonable request and subject to approval by the institutional review board.

Conflicts of Interest

JK, LK, and MB are employees of Autoscriber. Their affiliation with Autoscriber did not influence the study design, data collection and analysis, decision to publish, or preparation of the manuscript. The other authors, who are not affiliated with Autoscriber, contributed independently to this work, ensuring unbiased data interpretation and conclusions.

Multimedia Appendix 1

Extended Table 2: Descriptive statistics of the different methods and associated *P* values.

[[DOCX File, 16 KB - ai_v3i1e60020_app1.docx](#)]

Multimedia Appendix 2

Prompt provided to ChatGPT, version 4.0.

[[DOCX File, 13 KB - ai_v3i1e60020_app2.docx](#)]

Multimedia Appendix 3

Differences in PDQI-9 (Physician Documentation Quality Instrument) score between automatic and manual summaries per student and in time spent on manual and edited summaries per student.

[[DOCX File, 58 KB - ai_v3i1e60020_app3.docx](#)]

References

1. Shanafelt TD, West CP, Sinsky C, Trockel M, Tutty M, Satele DV, et al. Changes in burnout and satisfaction with work-life integration in physicians and the general US working population between 2011 and 2017. *Mayo Clin Proc* 2019;94(9):1681-1694 [FREE Full text] [doi: [10.1016/j.mayocp.2018.10.023](https://doi.org/10.1016/j.mayocp.2018.10.023)] [Medline: [30803733](https://pubmed.ncbi.nlm.nih.gov/30803733/)]
2. Taking Action Against Clinician Burnout: A Systems Approach to Professional Well-Being. Washington DC: The National Academies Press; 2019.
3. Arndt BG, Beasley JW, Watkinson MD, Temte JL, Tuan WJ, Sinsky CA, et al. Tethered to the EHR: primary care physician workload assessment using EHR event log data and time-motion observations. *Ann Fam Med* 2017;15(5):419-426 [FREE Full text] [doi: [10.1370/afm.2121](https://doi.org/10.1370/afm.2121)] [Medline: [28893811](https://pubmed.ncbi.nlm.nih.gov/28893811/)]
4. Sinsky C, Colligan L, Li L, Prgomet M, Reynolds S, Goeders L, et al. Allocation of physician time in ambulatory practice: a time and motion study in 4 specialties. *Ann Intern Med* 2016;165(11):753-760. [doi: [10.7326/M16-0961](https://doi.org/10.7326/M16-0961)] [Medline: [27595430](https://pubmed.ncbi.nlm.nih.gov/27595430/)]
5. Tai-Seale M, Olson CW, Li J, Chan AS, Morikawa C, Durbin M, et al. Electronic health record logs indicate that physicians split time evenly between seeing patients and desktop medicine. *Health Aff (Millwood)* 2017;36(4):655-662 [FREE Full text] [doi: [10.1377/hlthaff.2016.0811](https://doi.org/10.1377/hlthaff.2016.0811)] [Medline: [28373331](https://pubmed.ncbi.nlm.nih.gov/28373331/)]
6. Rao SK, Kimball AB, Lehrhoff SR, Hidrue MK, Colton DG, Ferris TG, et al. The impact of administrative burden on academic physicians: results of a hospital-wide physician survey. *Acad Med* 2017;92(2):237-243. [doi: [10.1097/ACM.0000000000001461](https://doi.org/10.1097/ACM.0000000000001461)] [Medline: [28121687](https://pubmed.ncbi.nlm.nih.gov/28121687/)]
7. Pelland KD, Baier RR, Gardner RL. "It's like texting at the dinner table": a qualitative analysis of the impact of electronic health records on patient-physician interaction in hospitals. *J Innov Health Inform* 2017;24(2):894 [FREE Full text] [doi: [10.14236/jhi.v24i2.894](https://doi.org/10.14236/jhi.v24i2.894)] [Medline: [28749316](https://pubmed.ncbi.nlm.nih.gov/28749316/)]
8. Gardner RL, Cooper E, Haskell J, Harris DA, Poplau S, Kroth PJ, et al. Physician stress and burnout: the impact of health information technology. *J Am Med Inform Assoc* 2019;26(2):106-114 [FREE Full text] [doi: [10.1093/jamia/ocy145](https://doi.org/10.1093/jamia/ocy145)] [Medline: [30517663](https://pubmed.ncbi.nlm.nih.gov/30517663/)]
9. Robertson SL, Robinson MD, Reid A. Electronic health record effects on work-life balance and burnout within the I3 population collaborative. *J Grad Med Educ* 2017;9(4):479-484 [FREE Full text] [doi: [10.4300/JGME-D-16-00123.1](https://doi.org/10.4300/JGME-D-16-00123.1)] [Medline: [28824762](https://pubmed.ncbi.nlm.nih.gov/28824762/)]
10. Coiera E, Kocaballi B, Halamka J, Laranjo L. The digital scribe. *NPJ Digit Med* 2018;1:58 [FREE Full text] [doi: [10.1038/s41746-018-0066-9](https://doi.org/10.1038/s41746-018-0066-9)] [Medline: [31304337](https://pubmed.ncbi.nlm.nih.gov/31304337/)]
11. Quiroz JC, Laranjo L, Kocaballi AB, Berkovsky S, Rezazadegan D, Coiera E. Challenges of developing a digital scribe to reduce clinical documentation burden. *NPJ Digit Med* 2019;2:114 [FREE Full text] [doi: [10.1038/s41746-019-0190-1](https://doi.org/10.1038/s41746-019-0190-1)] [Medline: [31799422](https://pubmed.ncbi.nlm.nih.gov/31799422/)]
12. van Buchem MM, Boosman H, Bauer MP, Kant IMJ, Cammel SA, Steyerberg EW. The digital scribe in clinical practice: a scoping review and research agenda. *NPJ Digit Med* 2021;4(1):57 [FREE Full text] [doi: [10.1038/s41746-021-00432-5](https://doi.org/10.1038/s41746-021-00432-5)] [Medline: [33772070](https://pubmed.ncbi.nlm.nih.gov/33772070/)]
13. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. *Nat Med* 2023;29(8):1930-1940. [doi: [10.1038/s41591-023-02448-8](https://doi.org/10.1038/s41591-023-02448-8)] [Medline: [37460753](https://pubmed.ncbi.nlm.nih.gov/37460753/)]
14. Giorgi J, Toma A, Xie R, Chen S, An K, Zheng G, et al. WangLab at MEDIQA-Chat 2023: clinical note generation from doctor-patient conversations using large language models. arXiv:2305.02220 2023. [doi: [10.18653/v1/2023.clinicalnlp-1.36](https://doi.org/10.18653/v1/2023.clinicalnlp-1.36)]

15. Abacha A, Yim W, Fan Y, Lin T. An empirical study of clinical note generation from doctor-patient encounters. 2023 Presented at: Proc 17th Conf Eur Chapter Assoc Comput Linguistics; 2023 Aug 08; Dubrovnik, Croatia p. 2291-2302. [doi: [10.18653/v1/2023.eacl-main.168](https://doi.org/10.18653/v1/2023.eacl-main.168)]
16. Stetson PD, Bakken S, Wrenn JO, Siegler EL. Assessing electronic note quality using the physician documentation quality instrument (PDQI-9). *Appl Clin Inform* 2012;3(2):164-174 [FREE Full text] [doi: [10.4338/aci-2011-11-ra-0070](https://doi.org/10.4338/aci-2011-11-ra-0070)] [Medline: [22577483](https://pubmed.ncbi.nlm.nih.gov/22577483/)]
17. Lin CY. ROUGE: a package for automatic evaluation of summaries. In: Text Summarization Branches Out Barcelona. Spain: Association for Computational Linguistics; 2004 Presented at: Text Summarization Branches Out Barcelona; 2004-07-25; Barcelona p. 74-81 URL: <https://aclanthology.org/W04-1013/>
18. Zhang T, Kishore V, Wu F, Weinberger KQ, Artzi Y. BERTScore: evaluating text generation with BERT. arXiv:1904.09675 2019. [doi: [10.5260/chara.21.2.8](https://doi.org/10.5260/chara.21.2.8)]
19. Lee FC, Chong WF, Chong P, Ooi SB. The emergency medicine department system: a study of the effects of computerization on the quality of medical records. *Eur J Emerg Med* 2001;8(2):107-115. [doi: [10.1097/00063110-200106000-00006](https://doi.org/10.1097/00063110-200106000-00006)] [Medline: [11436906](https://pubmed.ncbi.nlm.nih.gov/11436906/)]
20. Lai FW, Kant JA, Dombagolla MH, Hendarto A, Ugoni A, Taylor DM. Variables associated with completeness of medical record documentation in the emergency department. *Emerg Med Australas* 2019;31(4):632-638. [doi: [10.1111/1742-6723.13229](https://doi.org/10.1111/1742-6723.13229)] [Medline: [30690885](https://pubmed.ncbi.nlm.nih.gov/30690885/)]
21. Soto CM, Kleinman KP, Simon SR. Quality and correlates of medical record documentation in the ambulatory care setting. *BMC Health Serv Res* 2002;2(1):22 [FREE Full text] [doi: [10.1186/1472-6963-2-22](https://doi.org/10.1186/1472-6963-2-22)] [Medline: [12473161](https://pubmed.ncbi.nlm.nih.gov/12473161/)]
22. Guo B, Zhang X, Wang Z, Jiang M, Nie J, Ding Y, et al. How close is ChatGPT to human Experts? Comparison corpus, evaluation, and detection. arXiv:2301.07597 2023. [doi: [10.48550/arxiv.2301.07597](https://doi.org/10.48550/arxiv.2301.07597)]
23. Liao W, Liu Z, Dai H, Xu S, Wu Z, Zhang Y, et al. Differentiating ChatGPT-generated and human-written medical texts: quantitative study. *JMIR Med Educ* 2023;9:e48904 [FREE Full text] [doi: [10.2196/48904](https://doi.org/10.2196/48904)] [Medline: [38153785](https://pubmed.ncbi.nlm.nih.gov/38153785/)]
24. Ebbers T, Kool RB, Smeele LE, Dirven R, den Besten CA, Karssemakers LHE, et al. The impact of structured and standardized documentation on documentation quality; a multicenter, retrospective study. *J Med Syst* 2022;46(7):46 [FREE Full text] [doi: [10.1007/s10916-022-01837-9](https://doi.org/10.1007/s10916-022-01837-9)] [Medline: [35618978](https://pubmed.ncbi.nlm.nih.gov/35618978/)]
25. Elkbuli A, Godelman S, Miller A, Boneva D, Bernal E, Hai S, et al. Improved clinical documentation leads to superior reportable outcomes: an accurate representation of patient's clinical status. *Int J Surg* 2018;53:288-291 [FREE Full text] [doi: [10.1016/j.ijisu.2018.03.081](https://doi.org/10.1016/j.ijisu.2018.03.081)] [Medline: [29653245](https://pubmed.ncbi.nlm.nih.gov/29653245/)]
26. Reyes C, Greenbaum A, Porto C, Russell JC. Implementation of a clinical documentation improvement curriculum improves quality metrics and hospital charges in an academic surgery department. *J Am Coll Surg* 2017;224(3):301-309. [doi: [10.1016/j.jamcollsurg.2016.11.010](https://doi.org/10.1016/j.jamcollsurg.2016.11.010)] [Medline: [27919741](https://pubmed.ncbi.nlm.nih.gov/27919741/)]
27. Kittinger BJ, Matejicka A, Mahabir RC. Surgical precision in clinical documentation connects patient safety, quality of care, and reimbursement. *Perspect Health Inf Manag* 2016;13(Winter):1f [FREE Full text] [Medline: [26903784](https://pubmed.ncbi.nlm.nih.gov/26903784/)]
28. Akter M, Bansal N, Karmaker SK. Revisiting Automatic Evaluation of Extractive Summarization Task: Can We Do Better than ROUGE? In: Findings of the Association for Computational Linguistics: ACL 2022.: Association for Computational Linguistics; 2022 Presented at: Annual Meeting of the Association for Computational Linguistics; 2022-05-22; Dublin p. 1547-1560. [doi: [10.18653/v1/2022.findings-acl.122](https://doi.org/10.18653/v1/2022.findings-acl.122)]
29. Tu T, Azizi S, Driess D, Schaekermann M, Amin M, Chang P, et al. Towards generalist biomedical AI. arXiv:2401.05654 2024. [doi: [10.1056/aioa2300138](https://doi.org/10.1056/aioa2300138)]

Abbreviations

EHR: electronic health record

NLP: natural language processing

PDQI-9: Physician Documentation Quality Instrument

ROUGE: Recall-Oriented Understudy for Gisting Evaluation

Edited by G Luo; submitted 29.04.24; peer-reviewed by S Mao, U Sinha; comments to author 24.05.24; revised version received 12.07.24; accepted 19.07.24; published 23.09.24.

Please cite as:

van Buchem MM, Kant IMJ, King L, Kazmaier J, Steyerberg EW, Bauer MP

Impact of a Digital Scribe System on Clinical Documentation Time and Quality: Usability Study

JMIR AI 2024;3:e60020

URL: <https://ai.jmir.org/2024/1/e60020>

doi: [10.2196/60020](https://doi.org/10.2196/60020)

PMID: [39312397](https://pubmed.ncbi.nlm.nih.gov/39312397/)

©Marieke Meija van Buchem, Ilse M J Kant, Liza King, Jacqueline Kazmaier, Ewout W Steyerberg, Martijn P Bauer. Originally published in JMIR AI (<https://ai.jmir.org>), 23.09.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR AI, is properly cited. The complete bibliographic information, a link to the original publication on <https://www.ai.jmir.org/>, as well as this copyright and license information must be included.

Research Letter

What Is the Performance of ChatGPT in Determining the Gender of Individuals Based on Their First and Last Names?

Paul Sebo¹, MSc, MD

University Institute for Primary Care, University of Geneva, Geneva, Switzerland

Corresponding Author:

Paul Sebo, MSc, MD

University Institute for Primary Care

University of Geneva

Rue Michel-Servet 1

Geneva, 1211

Switzerland

Phone: 41 223794390

Email: paulsebo@hotmail.com

(JMIR AI 2024;3:e53656) doi:[10.2196/53656](https://doi.org/10.2196/53656)

KEYWORDS

accuracy; artificial intelligence; AI; ChatGPT; gender; gender detection tool; misclassification; name; performance; gender detection; gender detection tools; inequalities; language model; NamSor; Gender API; Switzerland; physicians; gender bias; disparities; gender disparities; gender gap

Introduction

Accurate determination of gender from names is vital for addressing gender-related disparities in medicine and promoting inclusivity. Gender detection tools (GDTs) offer efficient solutions, enabling large-scale demographic analysis [1-3] to improve data quality and inform targeted interventions. Indeed, they can process thousands of names simultaneously, saving time and resources. However, most of them charge for more than a certain number of requests per month. We recently compared the performance of 4 GDTs and showed that Gender API (Gender-API.com) and NamSor (NamSor Applied Onomastics) were accurate (misclassifications=1.5% and 2.0%, respectively; nonclassifications=0.3% and 0%, respectively) [4].

ChatGPT is a language model developed by OpenAI that is capable of generating human-like text and engaging in natural language conversations [5]. In medicine, ChatGPT can be employed for various purposes, such as answering patient queries and providing information on medical topics, making it a valuable resource for health care professionals and researchers seeking quick access to medical information and support in their work [6,7].

Given the increasing usefulness of GDTs in research, particularly for evaluating gender disparities in medicine, we assessed whether the performance of ChatGPT as a free GDT (version GPT-3.5) could approach that of Gender API and NamSor. We also compared ChatGPT-3.5 with the more advanced GPT-4 version. We hypothesized that ChatGPT, a

versatile language model not specifically trained for gender analysis, could achieve gender detection performance comparable to specialized tools and that ChatGPT-4 would perform no better than ChatGPT-3.5.

Methods

Database Selection and Data Collection

The methods used in this study are the same as those used in our primary study, which compared the performance of 4 GDTs [4]. We used a database of 6131 physicians practicing in Switzerland, a multilingual and multicultural country with 36% of physicians of foreign origin [4]. The sample consisted of 3085 women (50.3%) and 3046 men (49.7%), with gender determined by self-identification. We used nationalize.io to determine the origin of physicians' names (Table 1). A total of 88% of names were from French-, English-, Spanish-, Italian-, German-, or Portuguese-speaking countries or from another European country.

We asked ChatGPT-3.5 to determine the gender of 500 physicians at a time, after copying and pasting these lists of first and last names from the database. We ran the analysis twice and also examined ChatGPT-4 to check the "stability" of the responses [8]. The data were collected between September and November 2023.

We constructed a confusion matrix (Table 2): *ff* and *mm* correspond to correct classifications, *mf* and *fm* to misclassifications, and *fu* and *mu* to nonclassifications (ie, gender impossible to determine).

As in other studies [4,9], we calculated 4 performance metrics, namely “errorCoded” (the proportion of misclassifications and nonclassifications), “errorCodedWithoutNA” (the proportion of misclassifications), “naCoded” (the proportion of

nonclassifications), and “errorGenderBias” (the direction of bias in gender determination). We used Cohen κ to assess interrater agreement.

Table 1. Estimated origin of physicians’ names (N=6131 physicians).

Origin	Count ^a , n (%)
French-speaking country	1679 (32.2)
English-speaking country	751 (14.4)
Spanish-speaking country	404 (7.7)
Asian country ^b	344 (6.6)
Eastern European country	324 (6.2)
Italian-speaking country	288 (5.5)
Western European country ^b	272 (5.2)
Arabic-speaking country	259 (5.0)
German-speaking country	259 (5.0)
Northern European country ^b	220 (4.2)
Southern European country ^b	217 (4.2)
Portuguese-speaking country	198 (3.8)

^aThe total number of physicians does not add to 6131 because of missing values (no assignments for 916 physicians).

^bIf not already classified in another group (eg, in the Arabic-speaking country group for some Asian countries).

Table 2. Confusion matrix showing the 6 possible classification outcomes.

	Female (predicted)	Male (predicted)	Unknown (predicted)
Female (actual)	ff	fm	fu
Male (actual)	mf	mm	mu

Ethical Considerations

Since this study did not involve the collection of personal health-related data, it did not require ethical review per current Swiss law.

Results

Performance metrics showed high accuracy for ChatGPT-3.5 and ChatGPT-4 in both the first and second rounds (Table 3).

The number of misclassifications was low (proportion $\leq 1.5\%$) and there were no “nonclassifications.” As shown in Table 3, interrater agreement between the first and second rounds (for ChatGPT-3.5 and ChatGPT-4) and between ChatGPT-3.5 and ChatGPT-4 (for the first round) was “almost perfect” ($\kappa > 0.97$, all $P_s < .001$).

Table 3. Confusion matrix and performance metrics for ChatGPT-3.5 and ChatGPT-4 (N=6131 physicians).

	Classified as women, n (%)	Classified as men, n (%)	Unclassified, n (%)	Interrater agreement ^a	
				Cohen κ (95% CI)	<i>P</i> value
ChatGPT-3.5				0.9817 (0.9770-0.9865) ^b	<.001
First round^c					
Female physicians (n=3085)	3028 (98.2)	57 (1.8)	0 (0)		
Male physicians (n=3046)	18 (0.6)	3028 (99.4)	0 (0)		
Second round^d					
Female physicians (n=3085)	3030 (98.2)	55 (1.8)	0 (0)		
Male physicians (n=3046)	28 (0.9)	3018 (99.1)	0 (0)		
ChatGPT-4				0.9958 (0.9935-0.9981) ^b	<.001
First round^e					
Female physicians (n=3085)	3020 (97.9)	65 (2.1)	0 (0)		
Male physicians (n=3046)	27 (0.9)	3019 (99.1)	0 (0)		
Second round^f					
Female physicians (n=3085)	3020 (97.9)	65 (2.1)	0 (0)		
Male physicians (n=3046)	26 (0.9)	3020 (99.1)	0 (0)		

^aInterrater agreement between ChatGPT-3.5 and ChatGPT-4 (for the first round): Cohen κ =0.9768, 95% CI 0.9715-0.9822, *P*<.001.

^bInterrater agreement between the first and second rounds for each version.

^cPerformance metrics: errorCoded=0.01223, errorCodedWithoutNA=0.01223, naCoded=0, and errorGenderBias=-0.00636.

^dPerformance metrics: errorCoded=0.01354, errorCodedWithoutNA=0.01354, naCoded=0, and errorGenderBias=-0.00440.

^ePerformance metrics: errorCoded=0.01501, errorCodedWithoutNA=0.01501, naCoded=0, and errorGenderBias=-0.00620.

^fPerformance metrics: errorCoded=0.01484, errorCodedWithoutNA=0.01484, naCoded=0, and errorGenderBias=-0.00636.

Discussion

We used ChatGPT to determine the gender of 6131 physicians practicing in Switzerland and found that the proportion of misclassifications was $\leq 1.5\%$ for both versions. There were no nonclassifications and gender bias was negligible. Interrater agreement between ChatGPT-3.5 and ChatGPT-4 was “almost perfect.”

These results are relatively similar to those found in our primary study for Gender API and NamSor (errorCoded=0.0181 and 0.0202, errorCodedWithoutNA=0.0147 and 0.0202, naCoded=0.0034 and 0, errorGenderBias=-0.0072 and 0.0026) [4]. They are slightly better than those of another study published in 2018, which compared 5 GDTs, including Gender API and NamSor [9]. These results suggest that ChatGPT can

accurately determine the gender of individuals using their first and last names. The disadvantage of ChatGPT compared to Gender API and NamSor is that the database cannot be uploaded directly into ChatGPT (eg, as an Excel or CSV file).

Both ChatGPT-3.5 and ChatGPT-4 exhibit high accuracy in gender detection, with no significant superiority observed in ChatGPT-4 over ChatGPT-3.5. This underscores the robustness of ChatGPT in gender prediction across different versions. Our short study has 2 main limitations. Given the estimated origin of physicians' names, the results of the study can probably be generalized to most Western countries but not necessarily to Asian or Middle Eastern countries. GDTs are often less accurate with names from these countries [9,10]. In addition, GDTs oversimplify the concept of gender by dichotomizing individuals into male or female.

Data Availability

The data associated with this article are available in the Open Science Framework [11].

Conflicts of Interest

None declared.

References

1. Cevik M, Haque S, Manne-Goehler J, Kuppalli K, Sax PE, Majumder MS, et al. Gender disparities in coronavirus disease 2019 clinical trial leadership. *Clin Microbiol Infect* 2021 Jul;27(7):1007-1010 [FREE Full text] [doi: [10.1016/j.cmi.2020.12.025](https://doi.org/10.1016/j.cmi.2020.12.025)] [Medline: [33418021](https://pubmed.ncbi.nlm.nih.gov/33418021/)]
2. Sebo P, Clair C. Gender gap in authorship: a study of 44,000 articles published in 100 high-impact general medical journals. *Eur J Intern Med* 2022 Mar;97:103-105. [doi: [10.1016/j.ejim.2021.09.013](https://doi.org/10.1016/j.ejim.2021.09.013)] [Medline: [34598855](https://pubmed.ncbi.nlm.nih.gov/34598855/)]
3. Gottlieb M, Krzyzaniak SM, Mannix A, Parsons M, Mody S, Kalantari A, et al. Sex distribution of editorial board members among emergency medicine journals. *Ann Emerg Med* 2021 Jan;77(1):117-123. [doi: [10.1016/j.annemergmed.2020.03.027](https://doi.org/10.1016/j.annemergmed.2020.03.027)] [Medline: [32376090](https://pubmed.ncbi.nlm.nih.gov/32376090/)]
4. Sebo P. Performance of gender detection tools: a comparative study of name-to-gender inference services. *J Med Libr Assoc* 2021 Jul 01;109(3):414-421 [FREE Full text] [doi: [10.5195/jmla.2021.1185](https://doi.org/10.5195/jmla.2021.1185)] [Medline: [34629970](https://pubmed.ncbi.nlm.nih.gov/34629970/)]
5. Rao A, Pang M, Kim J, Kamineni M, Lie W, Prasad AK, et al. Assessing the utility of ChatGPT throughout the entire clinical workflow: development and usability study. *J Med Internet Res* 2023 Aug 22;25:e48659 [FREE Full text] [doi: [10.2196/48659](https://doi.org/10.2196/48659)] [Medline: [37606976](https://pubmed.ncbi.nlm.nih.gov/37606976/)]
6. Dave T, Athaluri SA, Singh S. ChatGPT in medicine: an overview of its applications, advantages, limitations, future prospects, and ethical considerations. *Front Artif Intell* 2023 May 4;6:1169595 [FREE Full text] [doi: [10.3389/frai.2023.1169595](https://doi.org/10.3389/frai.2023.1169595)] [Medline: [37215063](https://pubmed.ncbi.nlm.nih.gov/37215063/)]
7. Sallam M. ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. *Healthcare (Basel)* 2023 Mar 19;11(6):887 [FREE Full text] [doi: [10.3390/healthcare11060887](https://doi.org/10.3390/healthcare11060887)] [Medline: [36981544](https://pubmed.ncbi.nlm.nih.gov/36981544/)]
8. Zhu L, Mou W, Yang T, Chen R. ChatGPT can pass the AHA exams: open-ended questions outperform multiple-choice format. *Resuscitation* 2023 Jul;188:109783. [doi: [10.1016/j.resuscitation.2023.109783](https://doi.org/10.1016/j.resuscitation.2023.109783)] [Medline: [37349064](https://pubmed.ncbi.nlm.nih.gov/37349064/)]
9. Santamaría L, Mihaljević H. Comparison and benchmark of name-to-gender inference services. *PeerJ Comput Sci* 2018;4:e156 [FREE Full text] [doi: [10.7717/peerj-cs.156](https://doi.org/10.7717/peerj-cs.156)] [Medline: [33816809](https://pubmed.ncbi.nlm.nih.gov/33816809/)]
10. Sebo P. How accurate are gender detection tools in predicting the gender for Chinese names? A study with 20,000 given names in Pinyin format. *J Med Libr Assoc* 2022 Apr 01;110(2):205-211 [FREE Full text] [doi: [10.5195/jmla.2022.1289](https://doi.org/10.5195/jmla.2022.1289)] [Medline: [35440899](https://pubmed.ncbi.nlm.nih.gov/35440899/)]
11. What is the performance of ChatGPT in determining the gender of individuals based on their first and last names? Open Science Framework. 2023 Sep 27. URL: <https://osf.io/6nzd4/> [accessed 2024-03-08]

Abbreviations

GDT: gender detection tool

Edited by K El Emam, B Malin; submitted 14.10.23; peer-reviewed by ZA Teel, A Shamsi, L Zhu; comments to author 21.11.23; revised version received 26.11.23; accepted 02.03.24; published 13.03.24.

Please cite as:

Sebo P

What Is the Performance of ChatGPT in Determining the Gender of Individuals Based on Their First and Last Names?

JMIR AI 2024;3:e53656

URL: <https://ai.jmir.org/2024/1/e53656>

doi: [10.2196/53656](https://doi.org/10.2196/53656)

PMID:

©Paul Sebo. Originally published in JMIR AI (<https://ai.jmir.org>), 13.03.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR AI, is properly cited. The complete bibliographic information, a link to the original publication on <https://www.ai.jmir.org/>, as well as this copyright and license information must be included.

Original Paper

Machine Learning Methods Using Artificial Intelligence Deployed on Electronic Health Record Data for Identification and Referral of At-Risk Patients From Primary Care Physicians to Eye Care Specialists: Retrospective, Case-Controlled Study

Joshua A Young¹, MD; Chin-Wen Chang², PhD; Charles W Scales³, PhD; Saurabh V Menon⁴, BTech; Chantal E Holy⁵, MS, PhD; Caroline Adrienne Blackie⁶, MS, OD, PhD

¹Department of Ophthalmology, New York University School of Medicine, New York, NY, United States

²Data Science, Johnson & Johnson MedTech, Raritan, NJ, United States

³Medical and Scientific Operations, Johnson & Johnson Medtech, Vision, Jacksonville, FL, United States

⁴Mu Sigma Business Solutions Private Limited, Bangalore, India

⁵Epidemiology and Real-World Data Sciences, Johnson & Johnson MedTech, New Brunswick, NJ, United States

⁶Medical and Scientific Operations, Johnson & Johnson MedTech, Vision, Jacksonville, FL, United States

Corresponding Author:

Caroline Adrienne Blackie, MS, OD, PhD

Medical and Scientific Operations

Johnson & Johnson MedTech, Vision

7500 Centurion Parkway

Jacksonville, FL, 32256

United States

Phone: 1 9044331000

Email: cblackie@its.jnj.com

Abstract

Background: Identification and referral of at-risk patients from primary care practitioners (PCPs) to eye care professionals remain a challenge. Approximately 1.9 million Americans suffer from vision loss as a result of undiagnosed or untreated ophthalmic conditions. In ophthalmology, artificial intelligence (AI) is used to predict glaucoma progression, recognize diabetic retinopathy (DR), and classify ocular tumors; however, AI has not yet been used to triage primary care patients for ophthalmology referral.

Objective: This study aimed to build and compare machine learning (ML) methods, applicable to electronic health records (EHRs) of PCPs, capable of triaging patients for referral to eye care specialists.

Methods: Accessing the Optum deidentified EHR data set, 743,039 patients with 5 leading vision conditions (age-related macular degeneration [AMD], visually significant cataract, DR, glaucoma, or ocular surface disease [OSD]) were exact-matched on age and gender to 743,039 controls without eye conditions. Between 142 and 182 non-ophthalmic parameters per patient were input into 5 ML methods: generalized linear model, L1-regularized logistic regression, random forest, Extreme Gradient Boosting (XGBoost), and J48 decision tree. Model performance was compared for each pathology to select the most predictive algorithm. The area under the curve (AUC) was assessed for all algorithms for each outcome.

Results: XGBoost demonstrated the best performance, showing, respectively, a prediction accuracy and an AUC of 78.6% (95% CI 78.3%-78.9%) and 0.878 for visually significant cataract, 77.4% (95% CI 76.7%-78.1%) and 0.858 for exudative AMD, 79.2% (95% CI 78.8%-79.6%) and 0.879 for nonexudative AMD, 72.2% (95% CI 69.9%-74.5%) and 0.803 for OSD requiring medication, 70.8% (95% CI 70.5%-71.1%) and 0.785 for glaucoma, 85.0% (95% CI 84.2%-85.8%) and 0.924 for type 1 nonproliferative diabetic retinopathy (NPDR), 82.2% (95% CI 80.4%-84.0%) and 0.911 for type 1 proliferative diabetic retinopathy (PDR), 81.3% (95% CI 81.0%-81.6%) and 0.891 for type 2 NPDR, and 82.1% (95% CI 81.3%-82.9%) and 0.900 for type 2 PDR.

Conclusions: The 5 ML methods deployed were able to successfully identify patients with elevated odds ratios (ORs), thus capable of patient triage, for ocular pathology ranging from 2.4 (95% CI 2.4-2.5) for glaucoma to 5.7 (95% CI 5.0-6.4) for type 1 NPDR, with an average OR of 3.9. The application of these models could enable PCPs to better identify and triage patients at

risk for treatable ophthalmic pathology. Early identification of patients with unrecognized sight-threatening conditions may lead to earlier treatment and a reduced economic burden. More importantly, such triage may improve patients' lives.

(JMIR AI 2024;3:e48295) doi:[10.2196/48295](https://doi.org/10.2196/48295)

KEYWORDS

decision support for health professionals; tools, programs and algorithms; electronic health record; primary care; artificial intelligence; AI; prediction accuracy; triaging; AI model; eye care; ophthalmic

Introduction

In the United States alone, more than 93 million adults were at high risk for vision loss in 2017; however, only 56.9% visited an eye care professional annually, and only 59.8% received a dilated eye examination [1]. More than 4 million Americans suffer from uncorrectable vision impairment, and more than 1 million are blind; this number is predicted to more than double by 2050 to 9 million due to the increasing epidemics of diabetes and other chronic diseases and our rapidly aging US population [2]. The impact of poor eyesight is manifest in its potentiation of comorbidities, particularly in increasing the risk of disability in patients with cognitive impairment [3]. Early identification of patients with unrecognized sight-threatening conditions may lead to earlier treatment and a reduced economic burden. More importantly, such triage may improve patients' lives.

The identification and referral of patients at risk of vision loss from primary care practitioners (PCPs) to eye care professionals remains a challenge [4]. A 2010 study identified a number of barriers, including a lack of access to ophthalmic screening within the setting of the PCP's office [4]. Some regional efforts have been made to improve the efficiency of triage of patients at risk for glaucoma [5] and diabetic retinopathy (DR) [6]; however, existing initiatives triage patients on only a few demographic and comorbidity parameters, whereas many systemic associations have been identified for age-related macular degeneration (AMD), cataract, DR, glaucoma, and ocular surface disease (OSD) [7-16].

Artificial intelligence (AI) modeling techniques are becoming increasingly important in ophthalmology in particular and medicine in general [17-20]. In ophthalmology, AI is used to calculate intraocular lens (IOL) powers [21-23], predict glaucoma progression [24,25], recognize DR [26], and classify ocular tumors [27]. To the best of our knowledge, AI has not yet been used to triage primary care patients for ophthalmology referral. In this study, the development, validation, and testing of multiple predictive machine learning (ML) methods for 5 leading sight-threatening and treatable ocular pathologies (ie, AMD, visually significant cataract, DR, glaucoma, and OSD) that have the potential to be used by PCPs to triage patients, based on existing data in their electronic health records (EHRs), for referral to eye care specialists were reported.

Methods

AI Modeling

All AI techniques have in common the process of "training," the adjustment of importance (ie, weights) of attributes or intermediate values, based on a set of data referred to as a

training set. The model performance is then assessed against another set of data called the test set. Similar model performance on training and test sets demonstrates model generalizability. The advent of large clinical databases has made possible the construction and training of both ML and neural network AI models. To this end, a large commercial EHR database that includes demographic, diagnostic, and therapeutic data to create and curate an ophthalmologically focused data set from which predictive models of multiple eye diseases can be built was used. We chose to compare several different ML methods to create models that might be used by PCPs to triage patients for referral to an eye care specialist. The models thus created used non-ophthalmic clinical and demographic data to assess relative risk scores for AMD, cataract, DR, glaucoma, and OSD.

Data Source

This retrospective, case-controlled study used data from the Optum deidentified EHR data set. EHRs provide efficient access to detailed patient-level longitudinal data that represent integral components of clinical care that may not necessarily be available through other retrospective database sources, such as administrative claims databases or patient registries [28,29]. The Optum EHR data set consists of data primarily from the United States and represents the clinical information of more than 80 million patients, including at least 7 million patients in each US census region from May 2000 to December 2019. Data from multiple EHR platforms, including Cerner, Epic, GE, and McKesson, are analyzed by Optum by means of natural language processing (NLP) to extract information about patient demographics, enrollment, diagnoses, biometrics, laboratory results, procedures, and medications [30]. The data set draws upon a network of more than 140,000 providers at more than 700 hospitals and 7000 clinics.

Ethical Considerations

The use of the Optum EHR data set was reviewed by the New England Institutional Review Board (IRB) and was determined to be exempt from broad IRB approval as this research project did not involve human subject research.

Outcome Measures

This study sought to predict the diagnosis of 5 major eye pathologies: AMD, cataract, DR, glaucoma, and OSD. The classification of AMD was based on the *International Classification of Diseases, 10th Revision* (ICD-10) codes and subdivided into nonexudative (H35.31%) and exudative (H35.32%) groups, in which "%" represents a wildcard. The classification of cataract required a more restrictive definition than simply H25%. Since no ICD-10 code distinguishes visually significant cataracts from those of lesser impact, we chose to use cataract surgery as a surrogate for visually significant

cataract. For this study, cataract was defined by the cataract surgery Current Procedural Terminology (CPT) codes of 66982 and 66984 rather than by ICD-10. The classification of DR was based on the data set ICD-10 codes and subdivided into type 1 nonproliferative diabetic retinopathy (NPDR; H10.31%-H10.34%), type 1 proliferative diabetic retinopathy (PDR; H10.35%), type 2 NPDR (H11.31%-H11.34%), and type 2 PDR (H11.35%). Glaucoma was defined by the presence of 1 or more of 3 criteria: an ICD-10 code of H40.1% (open-angle glaucoma), the prescription of glaucoma medication, or the presence of a CPT code indicating glaucoma surgery. This

definition was developed to capture not only patients with a recorded diagnosis of glaucoma but also those patients being treated for glaucoma or high-risk ocular hypertension for whom the diagnosis of glaucoma was not recorded in the data set. Similar to cataract, OSD required narrower criteria than simply H04.1% and H02.88% since these codes do not distinguish OSD requiring treatment from more mild presentations. For this study, OSD was defined rather restrictively as patients receiving cyclosporine ophthalmic emulsion 0.05%, cyclosporine ophthalmic solution 0.09%, or lifitegrast ophthalmic solution 5% (see [Tables 1](#) and [2](#)).

Table 1. Listed medications for glaucoma.

Type of medication	Examples
Beta blockers	Levobunolol (Betagan, Akbeta), timolol (Timoptic, Betimal, Istalol), carteolol (Ocupress), metipranolol (Optipranolol), timolol gel (Timoptic Xe), betaxolol (Betoptic, Betoptic S)
Alpha agonists	Apraclonidine (Iopidine), brimonidine (Alphagan, Alphagan P), dipivefrin (Propine)
Carbonic anhydrase inhibitors	Dorzolamide (Trusopt), brinzolamide (Azopt)
Prostaglandin analogs	Latanoprost (Xalatan), bimatoprost 0.01% (Lumigan), travoprost (Travatan Z), tafluprost (Zioptan), latanoprostene bunod (Vyzulta)
Prostaglandin analogs (combined medications)	Dorzolamide/timolol (Cosopt and Cospot Pf), brimonidine/timolol (Combigan), brinzolamide/brimonidine (Simbrinza), netarsudil/latanoprost (Rocklatan)
Rho kinase inhibitors	Netarsudil (Rhopressa)

Table 2. Listed procedures for glaucoma.

ICD-10 ^a code	Description
0191T	Insertion of anterior segment aqueous drainage device, without extraocular reservoir, internal approach, into the trabecular meshwork; initial insertion
0253T	Insertion of anterior segment aqueous drainage device, without extraocular reservoir, internal approach, into the suprachoroidal space
0376T	Insertion of anterior segment aqueous drainage device, without extraocular reservoir, internal approach, into the trabecular meshwork; each additional device insertion (list separately in addition to code for primary procedure)
0449T	Insertion of aqueous drainage device, without extraocular reservoir, internal approach, into the subconjunctival space; initial device
0450T	Insertion of aqueous drainage device, without extraocular reservoir, internal approach, into the subconjunctival space; each additional device (list separately in addition to code for primary procedure)
0474T	Insertion of anterior segment aqueous drainage device, with creation of intraocular reservoir, internal approach, into the supraciliary space
65820	Goniotomy
65855	Trabeculoplasty laser
66174	Transluminal dilation of aqueous outflow canal; without retention of device or stent
66175	Transluminal dilation of aqueous outflow canal; with retention of device or stent
66179	Aqueous shunt to extraocular equatorial plate reservoir, external approach; without graft
66180	Aqueous shunt to extraocular equatorial plate reservoir, external approach; with graft
66183	Insertion of anterior segment aqueous drainage device, without extraocular reservoir, external approach
66184	Revision of aqueous shunt to extraocular equatorial plate reservoir; without graft
66185	Revision of aqueous shunt to extraocular equatorial plate reservoir; with graft
66710	ciliary body destruction by cyclophotocoagulation, trans-scleral approach
66711	ciliary body destruction by cyclophotocoagulation, endoscopic approach (endoscopic cyclophotocoagulation)

^aICD-10: International Classification of Diseases, 10th Revision.

Creation of Patient Cohorts

Five distinct cohorts (ocular cohorts) of patients (AMD $n=294,739$, cataract $n=1,191,492$, DR $n=348,056$, glaucoma $n=843,560$, and OSD $n=660,218$) were selected from the Optum EHR data set based on the aforementioned code definitions from October 2015 onward (to limit the analysis to the start of the ICD-10 coding system in the United States). The inclusion criteria were as follows: patients with diagnosis codes such as H3530% / H3531% / H3532% , H25% , E083%/E093%/E103%/E113%/E133% , H40% , or H041%/H0288% and EHRs with an ICD-10 diagnosis code type. Patients were excluded if they had an unknown birth year, were younger than 15 years, had less than 60 days of continuous enrollment in the database prior to their diagnosis, had a gender labeled as unknown, or had undergone a cataract-related procedure or diagnosis at baseline or not undergone a cataract-related procedure and diagnosis in the follow-up. Patients with multiple conditions (eg, glaucoma and OSD) were identified in both the glaucoma and OSD cohorts. For each patient, demographic information, complete clinical and drug use information, and comorbidities were identified. [Multimedia Appendix 1](#) presents the patient inclusion and exclusion criteria and attrition data. All patients with the diagnoses present in the database during the specified inclusion period were considered for inclusion. Finally, the patients were segregated into subsets based on the AMD subtype or the DR subtype. In addition, only those patients who had open-angle glaucoma, had consumed a glaucoma-related medication, had undergone a glaucoma-related procedure in the follow-up, or had consumed dry eye and meibomian gland dysfunction (DEMGD)-related medications in the follow-up were retained. The final cohorts were as follows: exudative AMD $n=32,072$ (10.9%), nonexudative AMD $n=114,839$ (39%), cataract $n=197,570$ (16.6%), type I NPDR $n=20,654$ (5.9%), type I PDR $n=4465$ (1.3%), type II NPDR $n=155,927$ (44.8%), type II PDR $n=21,032$ (6%), glaucoma $n=192,727$ (22.8%), and OSD $n=3720$ (0.6%).

For each of the 5 cohorts, a control population was created from the pool of patients without ocular conditions. The control populations were matched 1:1 to each ocular cohort using exact matching on age and gender. A total of 743,039 patients with AMD, visually significant cataract, DR, glaucoma, or OSD were available in the Optum deidentified EHR data set, so these were exact-matched on age and gender to 743,039 controls without eye conditions.

Machine Learning

Several distinct ML approaches were followed to model the outcomes described earlier. These included the generalized linear model (GLM) [31], L1-regularized logistic regression (L1-LR) [32], random forest (RF) [33], Extreme Gradient Boosting (XGBoost) [34], and J48 decision tree (DT) [35].

Data Preprocessing

The data set consisted of 380 attributes, including demographic information, diagnoses, biometrics, laboratory results, procedures, and medications. Since some of these attributes, particularly some of the laboratory tests, were only sparsely represented, the data were pruned to remove attributes (ie,

“features” in ML) with more than 20% missing values. Missing values were imputed with medians for continuous variables (eg, BMI), with a “Missing” group for categorical variables (eg, smoke or alcohol usage), and with the most frequent value for binary variables (eg, levels of lab test results). Winsorization of the data was performed to remove outliers and replace these with 0.1 and 99.9 percentile values. Further feature engineering was performed to remove or combine highly correlated features, such as “rheumatoid arthritis/collagen vascular disease” and its highly correlated cognate “connective tissue disease.” These feature engineering steps were performed individually for each case-controlled data set of each subpathology. The resultant data sets exhibited between 142 and 182 features after the above-described culling. The feature exclusion data sets for each of the 9 subpathologies were modeled using each of 5 distinct modeling strategies to produce a total of 45 individual ML models. These 45 models were produced and compared in a competitive fashion to identify the single-best model for each pathology.

Model Strategies

Logistic regression without regularization (LR), L1-LR, RF, and XGBoost models were performed in Python (3.8.5) using the Scikit-learn (0.23.2) and XGBoost (1.2.0) libraries. Next, 80% of the data were used for training, and 20% of the data were used for testing with 5-fold cross-validation. A grid search was used to optimize hyperparameters. For L1-LR, the regularization strength C was tuned. In the RF algorithm, the space of the number of trees and the maximum depth of each tree combination were searched. The hyperparameter tuning for XGBoost included the learning rate and the maximum depth of each tree. The ML modeling pipeline was established, and information of missing values fit and learned from the training data was applied to the test data set to avoid information leakage. J48 DT modeling, a Java-based implementation of the C4 tree, was performed in the WEKA ML workbench (University of Waikato). Finally, 10-fold cross-validation was used with an initial leaf size of 2% of the data set. The area under the curve (AUC) was assessed for all algorithms for each outcome to measure the overall performance of the binary classification models.

Results

Cohort Details

The demographic information of each cohort is shown in [Table 3](#). Briefly, the total populations for modeling, for each cohort, varied in size from 7440 to 395,140. Populations were mostly female for AMD, cataract, glaucoma, and OSD requiring medications, and the average age ranged from 51 to 80 years.

The performance of different ML strategies varied as well ([Figures 1](#) and [2](#) and [Table 4](#)), but in all cases, XGBoost demonstrated the best performance, showing, respectively, a prediction accuracy and an AUC of 78.6% (95% CI 78.3%-78.9%) and 0.878 for visually significant cataract, 77.4% (95% CI 76.7%-78.1%) and 0.858 for exudative AMD, 79.2% (95% CI 78.8%-79.6%) and 0.879 for nonexudative AMD, 72.2% (95% CI 69.9%-74.5%) and 0.803 for OSD requiring medication, 70.8% (95% CI 70.5%-71.1%) and 0.785 for

glaucoma, 85.0% (95% CI 84.2%-85.8%) and 0.924 for type 1 NPDR, 82.2% (95% CI 80.4%-84.0%) and 0.911 for type 1 PDR, 81.3% (95% CI 81.0%-81.6%) and 0.891 for type 2 NPDR, and 82.1% (95% CI 81.3%-82.9%) and 0.900 for type 2 PDR (Table 4). XGBoost identified several clinical attributes that were important for diagnosis prediction (Figure 3).

The top-performing models identified the following clinical and demographic features that were primarily contributing to the predictions for each pathology (Figure 3; continuous measures showed positive associations):

- Exudative AMD diagnosis prediction was associated, in order of importance, with average household income, percentage college education, geographical division (Middle Atlantic, East North Central, East South Central, New England, South Atlantic/West South Central, Mountain, West North Central, Pacific, other/unknown), the BMI, and the Elixhauser score (comorbidity index).
- Nonexudative AMD demonstrated similar associations. In order of importance, these were average household income, percentage college education, region (Northeast, Midwest, South, West, other/unknown), smoking, and the Elixhauser score.
- Glaucoma clinical associations, in order of importance, included average household income, percentage college education, adrenal or androgen use, the BMI, and race.

- Cataract clinical associations, in order of importance, included average household income, percentage college education, region, the BMI, and smoking.
- OSD associations, in order of importance, included average household income, percentage college education, geographical division, rheumatoid arthritis and connective tissue disease, and region.
- DR associations varied over different subpathologies but generally included the Elixhauser score, high serum glucose, the BMI, hypertension, chronic pulmonary disease, depression, cardiac arrhythmia, and obesity.

Performance in predicting the presence of pathology ranged from 71% in the case of glaucoma to 87% in the case of type 1 PDR, with an average performance of 80% across all groups. Since the intent was to identify at-risk patients, these performance values were used to determine disease odds ratios (ORs) according to the method described by Hogue et al [36].

Applying this to each of the models provided a clinically useful measure. The models identified patients with elevated ORs of the prevalence of pathology from 2.4 in the case of glaucoma to 5.7 in the case of type I NPDR, with an average OR of 3.9 (Table 5).

Table 3. Demographic information of each cohort with ocular disease. For each cohort, a control (age- and gender-matched) population of similar size was generated, without the condition of interest.

Characteristic	Exudative AMD ^a (n=32,072)	Nonexudative AMD (n=114,839)	Cataract (n=197,570)	OSD ^b requiring medication (n=3720)	Glaucoma (n=192,727)	Type I NPDR ^c (n=20,654)	Type I PDR ^d (n=4465)	Type II NPDR (n=155,927)	Type II PDR (n=21,032)
Age (years), mean (SD)	79.8 (10.4)	77.1 (10.7)	69.7 (9.9)	68.3 (14.0)	72.4 (13.3)	51.5 (16.0)	52.1 (14.6)	64.4 (12.9)	61.6 (12.7)
Gender (female), n (%)	19,885 (62.0)	70,971 (61.8)	115,183 (58.3)	3050 (82.0)	108,698 (56.4)	10,203 (49.4)	2170 (48.6)	77,028 (49.4)	10,032 (47.7)
Race, n (%)									
Asian	353 (1.1)	1608 (1.4)	3951 (2.0)	52 (1.4)	3662 (1.9)	186 (0.9)	31 (0.7)	4054 (2.6)	484 (2.3)
Black	374 (2.1)	2756 (2.4)	13,632 (6.9)	272 (7.3)	30,065 (15.6)	2231 (10.8)	545 (12.2)	24,948 (16.0)	3912 (18.6)
White	27,903 (87.0)	97,843 (85.2)	160,229 (81.1)	3281 (88.2)	139,342 (72.3)	16,337 (79.1)	3393 (76.0)	106,342 (68.2)	13,166 (62.6)
Unknown	3143 (9.8)	12,632 (11.0)	23,511 (11.9)	112 (3.0)	19,658 (10.2)	1900 (9.2)	500 (11.2)	20,582 (13.2)	3449 (16.4)
Ethnicity, n (%)									
Hispanic	513 (1.6)	2067 (1.8)	5927 (3.0)	86 (2.3)	7516 (3.9)	888 (4.3)	223 (5.0)	13,722 (8.8)	2608 (12.4)
Non-Hispanic	27,774 (86.6)	96,465 (84.0)	168,132 (85.1)	3553 (95.5)	164,589 (85.4)	17,804 (86.2)	3764 (84.3)	124,118 (79.6)	15,900 (75.6)
Unknown	3784 (11.8)	16,307 (14.2)	23,511 (11.9)	82 (2.2)	20,622 (10.7)	1962 (9.5)	478 (10.7)	18,088 (11.6)	2524 (12.0)
Education (college educated), n (%)	7761 (24.2)	27,906 (24.3)	47,614 (24.1)	868 (23.2)	47,411 (24.6)	4936 (23.9)	1058 (23.7)	37,111 (23.8)	4943 (23.5)
Size of control population, n	32,072	114,839	197,570	3720	192,727	20,654	4465	155,927	21,032
Total population for modeling (cohort+control), n	64,144	229,678	395,140	7440	385,454	41,308	8930	311,854	42,064

^aAMD: age-related macular degeneration.

^bOSD: ocular surface disease.

^cNPDR: nonproliferative diabetic retinopathy.

^dPDR: proliferative diabetic retinopathy.

Figure 1. Model accuracy by pathology degeneration; AUC = area under the curve; CI = confidence interval; J48 = Decision tree; LR = Logistic Regression without regularization; LR-L1 = L1-regularized logistic regression; NPDR = non-proliferative diabetic retinopathy; OSD = ocular surface disease; PDR = proliferative diabetic retinopathy; XGB = XGBoost.

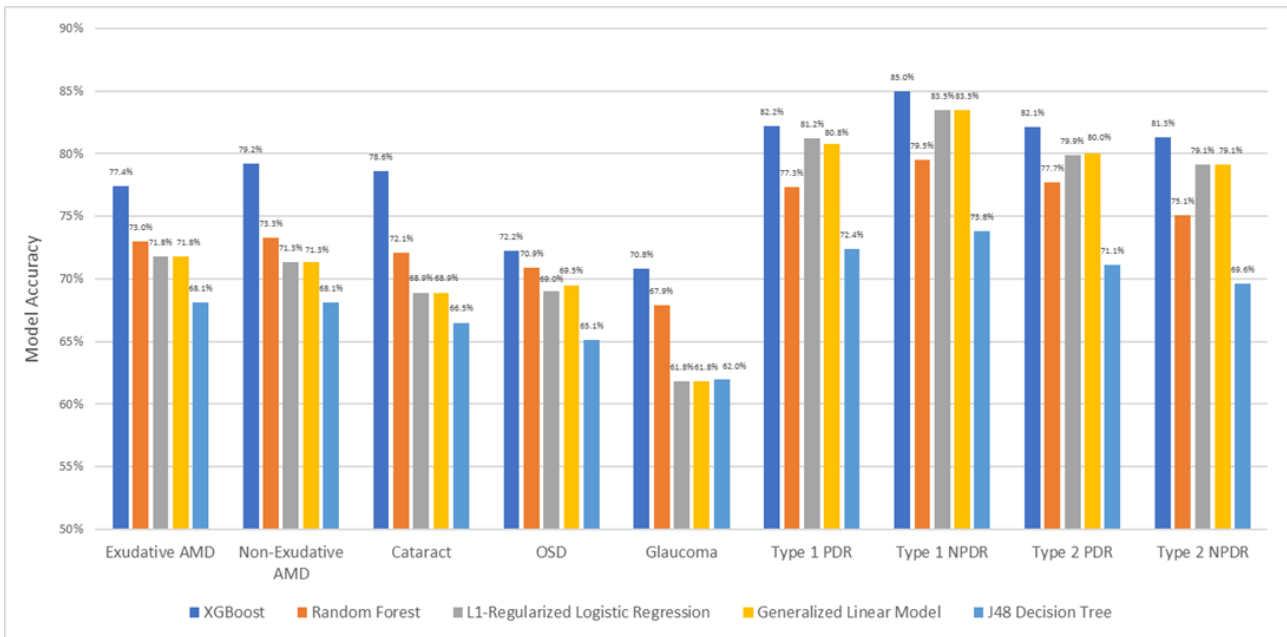


Figure 2. Receiver operating characteristic (ROC) curves illustrating the diagnostic ability of the models for the 9 pathologies. amd: age-related macular degeneration; auc: area under the curve; demgd: dry eye and meibomian gland dysfunction; j48: decision tree; l1: L1-regularized logistic regression; lr: logistic regression without regularization; npdr: nonproliferative diabetic retinopathy; pdr: proliferative diabetic retinopathy; rf: random forest; xgb: Extreme Gradient Boosting.

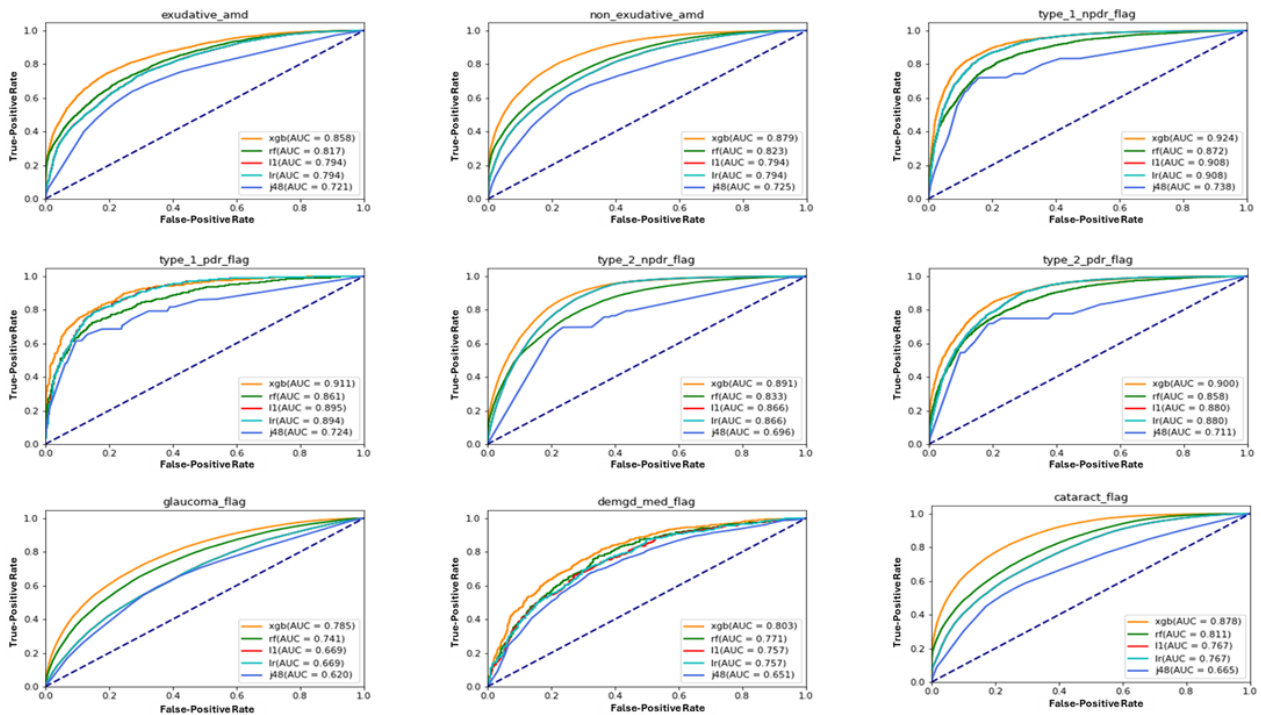


Table 4. Model accuracy, AUC^a, sensitivity, and specificity.

Outcome and algorithms	Accuracy (95% CI)	AUC (95% CI)	Sensitivity	Specificity
Cataract				
XGBoost ^b	78.6% (78.3%-78.9%)	0.878 (0.875-0.880)	0.796	0.776
RF ^c	72.1% (71.8%-72.4%)	0.811 (0.808-0.814)	0.749	0.693
LR-L1 ^d	68.9% (68.6%-69.2%)	0.767 (0.764-0.771)	0.683	0.695
LR ^e	68.9% (68.6%-69.2%)	0.767 (0.764-0.771)	0.683	0.695
J48 DT ^f	66.5% (N/A ^g)	0.710 (N/A)	0.702	0.628
Exudative AMD^h				
XGBoost	77.4% (76.7%-78.1%)	0.858 (0.851-0.863)	0.769	0.778
RF	73.0% (72.2%-73.8%)	0.817 (0.810-0.825)	0.745	0.715
LR-L1	71.8% (71.0%-72.6%)	0.794 (0.786-0.802)	0.716	0.720
LR	71.8% (71.0%-72.6%)	0.794 (0.786-0.801)	0.717	0.720
J48 DT	68.1% (N/A)	0.721 (N/A)	0.707	0.660
Nonexudative AMD				
XGBoost	79.2% (78.8%-79.6%)	0.879 (0.876-0.882)	0.801	0.783
RF	73.3% (72.9%-73.7%)	0.823 (0.820-0.827)	0.768	0.698
LR-L1	71.3% (70.9%-71.7%)	0.794 (0.790-0.798)	0.729	0.697
LR	71.3% (70.9%-71.7%)	0.794 (0.790-0.798)	0.727	0.700
J48 DT	68.1% (N/A)	0.725 (N/A)	0.741	0.622
OSDⁱ				
XGBoost	72.2% (69.9%-74.5%)	0.803 (0.780-0.824)	0.708	0.735
RF	70.9% (68.6%-73.2%)	0.771 (0.747-0.795)	0.749	0.669
LR-L1	69.0% (66.7%-71.3%)	0.757 (0.732-0.782)	0.691	0.688
LR	69.5% (67.2%-71.8%)	0.757 (0.733-0.782)	0.688	0.702
J48 DT	65.1% (N/A)	0.702 (N/A)	0.675	0.628
Glaucoma				
XGBoost	70.8% (70.5%-71.1%)	0.785 (0.782-0.788)	0.689	0.728
RF	67.9% (67.6%-68.2%)	0.741 (0.738-0.745)	0.656	0.702
LR-L1	61.8% (61.5%-62.1%)	0.669 (0.665-0.673)	0.622	0.614
LR	61.8% (61.5%-62.1%)	0.669 (0.665-0.673)	0.619	0.617
J48 DT	62.0% (N/A)	0.647 (N/A)	0.647	0.593
Type I NPDR^j				
XGBoost	85.0% (84.2%-85.8%)	0.924 (0.919-0.930)	0.850	0.850
RF	79.5% (78.6%-80.4%)	0.872 (0.864-0.879)	0.799	0.790
LR-L1	83.5% (82.7%-84.3%)	0.908 (0.902-0.915)	0.847	0.824
LR	83.5% (82.7%-84.3%)	0.908 (0.902-0.915)	0.847	0.824
J48 DT	73.8% (N/A)	0.796 (N/A)	0.756	0.721
Type I PDR^k				
XGBoost	82.2% (80.4%-84.0%)	0.911 (0.897-0.924)	0.816	0.828
RF	77.3% (75.4%-79.2%)	0.861 (0.846-0.878)	0.802	0.744

Outcome and algorithms	Accuracy (95% CI)	AUC (95% CI)	Sensitivity	Specificity
LR-L1	81.2% (79.4%-83.0%)	0.895 (0.881-0.910)	0.847	0.777
LR	80.8% (79.0%-82.6%)	0.894 (0.880-0.910)	0.829	0.787
J48 DT	72.4% (N/A)	0.804 (N/A)	0.761	0.686
Type II NPDR				
XGBoost	81.3% (81.0%-81.6%)	0.891 (0.888-0.893)	0.845	0.782
RF	75.1% (74.8%-75.4%)	0.833 (0.830-0.836)	0.751	0.752
LR-L1	79.1% (78.8%-79.4%)	0.866 (0.863-0.869)	0.843	0.739
LR	79.1% (78.8%-79.4%)	0.866 (0.863-0.869)	0.844	0.739
J48 DT	69.6% (N/A)	0.742 (N/A)	0.635	0.757
Type II PDR				
XGBoost	82.1% (81.3%-82.9%)	0.900 (0.893-0.907)	0.841	0.801
RF	77.7% (76.8%-78.6%)	0.858 (0.850-0.865)	0.763	0.790
LR-L1	79.9% (79.0%-80.8%)	0.880 (0.873-0.887)	0.834	0.763
LR	80.0% (79.1%-80.9%)	0.880 (0.873-0.887)	0.847	0.753
J48 DT	71.1% (N/A)	0.774 (N/A)	0.674	0.748

^aAUC: area under the curve.

^bXGBoost: Extreme Gradient Boosting.

^cRF: random forest.

^dL1-LR: L1-regularized logistic regression.

^eLR: logistic regression without regularization.

^fDT: decision tree.

^gN/A: not applicable.

^hAMD: age-related macular degeneration.

ⁱOSD: ocular surface disease.

^jNPDR: nonproliferative diabetic retinopathy.

^kPDR: proliferative diabetic retinopathy.

Figure 3. Clinical features primarily contributing to the predictions for each pathology. amd: age-related macular degeneration; demgd: dry eye and meibomian gland dysfunction; hh: household; npdr: nonproliferative diabetic retinopathy; pct: percentage; pdr: proliferative diabetic retinopathy; xgb: Extreme Gradient Boosting.

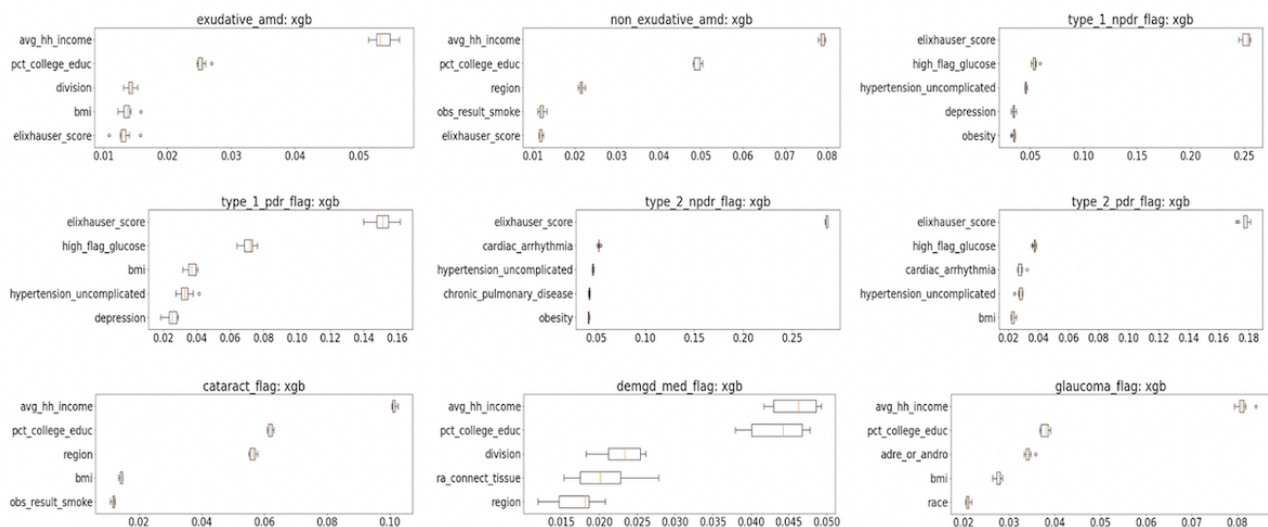


Table 5. Model accuracy and ORs^a by pathology.

Pathology	Model accuracy, %	OR (95% CI)
Exudative AMD ^b	77	3.4 (3.2-3.7)
Nonexudative AMD	79	3.8 (3.6-4.0)
Cataract	79	3.7 (3.6-3.8)
OSD ^c	72	2.6 (2.1-3.3)
Glaucoma	71	2.4 (2.4-2.5)
Type I PDR ^d	82	4.6 (3.6-5.9)
Type I NPDR ^e	85	5.7 (5.0-6.4)
Type II PDR	82	4.6 (4.1-5.1)
Type II NPDR	81	4.3 (4.2-4.5)

^aOR: odds ratio.

^bAMD: age-related macular degeneration.

^cOSD: ocular surface disease.

^dPDR: proliferative diabetic retinopathy.

^eNPDR: nonproliferative diabetic retinopathy.

Discussion

Principal Findings

A major challenge of current deep learning (DL) models is that their training requires a large amount of data because insufficient data may decrease the performance of DL models [37]. The original EHR data pool for this study comprised more than 80 million patients, one of the largest AI projects of its kind in ophthalmology. The final study populations totaled 1,486,078 patients, 50% of whom were controls. In addition to the substantial patient population, this study examined 9 subpathologies using 5 different analytical modeling approaches to identify the most predictive model for each pathology.

The goal of this effort was to create a digital health tool to identify patients at higher risk for the presence of ophthalmic pathology and to do this based solely on the sort of non-ophthalmic data to which a PCP would have access. The authors do not propose to either make definitive ophthalmic diagnoses or predict the development of future pathology. Rather, this work seeks to identify patients whose clinical and demographic context is associated with the presence of AMD, cataract, clinically significant DR, glaucoma, or OSD of a magnitude requiring pharmacological therapy. The creation, demonstration, and real-world validation (within a clinical setting) of a deployable digital tool will be the next step of this project.

The application of such a model in the clinical setting would allow a PCP to identify patients nearly 4 times more likely to have ophthalmic pathology. Such a tool would bring a substantial benefit in the triage and referral of at-risk patients to eye care professionals.

Data and Outcome Engineering

These data consist of diagnostic and procedure codes; biometric data, such as the BMI and vital signs; demographic information,

including socioeconomic and geographical information; laboratory results; and medications prescribed. This information does not include the physician notes that might provide a rationale for the diagnoses recorded. Indeed, since only a limited number of diagnoses may be listed on a claim, it is possible that some extant diagnoses may have gone unrecorded. However, diagnoses like cataract and OSD may be overrepresented since the ICD-10 taxonomy does not distinguish between clinically significant cataract and OSD from cases in which these pathologies are subclinical. Indeed, it would be of little clinical utility to build an AI model that detects subclinical cataracts.

Ours is not the first study to be faced with the challenge of identifying clinically relevant diagnoses from large data sets. A 2018 study [38] investigated the precision of ICD-10 codes for patients with uveitis and found that 13 of 27 uveitides were imprecisely defined and that multiple codes were used to describe the same pathology. A 2020 study of ocular pathology in patients with stroke [39] noted fewer patients with glaucoma than anticipated and attributed this to the lack of ophthalmology clinic data. The authors noted that patients may be on glaucoma medications without a concurrent ICD-10 code recorded for glaucoma, suggesting that a diagnosis of glaucoma may have been recorded in the patients' medical records before incorporation into the data set. The authors sought, therefore, to define the glaucoma cohort as those patients who met 1 or more of 3 criteria: an ICD-10 code of H40.1% (open-angle glaucoma), the prescription of glaucoma medication, or the presence of a CPT code indicating glaucoma surgery (see [Tables 1 and 2](#)). This definition was developed to both detect glaucoma patients without glaucoma ICD-10 codes and to exclude patients inappropriately labeled as glaucoma by ICD-10. This definition resulted in a substantial winnowing of the glaucoma cohort from 1,368,700, 50% of whom were controls, to 385,514 patients.

The authors took a similar approach to the cataract and OSD study populations. Cataract and OSD are among the most frequently recorded diagnoses on claims [40]. Cataract, in

particular, is nearly ubiquitous in elderly patients and was the most common ophthalmic ICD-10 diagnosis of those examined here. Since only a subset of these patients require cataract surgery, the detection of cataract alone is not clinically useful. ICD-10 coding does not distinguish between cataracts requiring surgery and those that do not. However, CPT coding, in a sense does make this distinction. Therefore, CPT codes of 66984 (cataract extraction with intraocular lens) and 66982 (complex cataract extraction) were chosen as the criteria for clinically significant cataracts. This narrowing of the inclusion criteria reduced our cataract study population from 2,087,836, 50% of whom were controls, to 395,140 patients. OSD coding is even more problematic. A large number of ICD-10 codes are available, and clinical significance is difficult to establish. Our initial cohort of OSD patients and controls totaled 1,182,912 patients. To model the clinical context associated with OSD, a restrictive criterion was chosen: the prescription of topical cyclosporine or lifitegrast. This greatly reduced the OSD population to 7440 patients, but this ensured the final population represented patients with clinically meaningful disease. No outcome engineering measures were applied to the AMD groups or to the DR groups, each of which was defined by its corresponding ICD-10 code.

In addition, PDR and NPDR could have been combined into 1 group since the referring physician probably would not care about what sort of DR the patient has. However, the NPDR group is so much larger than the PDR group that the authors do not expect that the segmentation is detrimental.

Clinical and Demographic Attributes and Feature Engineering

The initial data set included a large number of attributes or “features” (in the language of ML), totaling 380 individual parameters. To produce models that would not be burdensome for the clinician to use, the authors sought to reduce the number of attributes required by each model. This reduction and modification of model parameters is referred to as “feature engineering.” For a feature to be included in the final model, several criteria needed to be met. The feature must play a significant role in the model’s outcome. It is self-evident that features that do not contribute substantially to a model may be discarded with little impact on model performance. In the case of the XGBoost models, parameter optimization was performed by the grid search algorithm [41]. The second feature inclusion criterion was noncorrelation with other features. In some cases, such as between weight and the BMI, the correlation is evident. However, the correlation between other clinical features only becomes clear on analysis. The issue of feature correlation highlights a difference between AI and traditional risk analysis studies. When studied individually, certain attributes, such as obesity and socioeconomic status, may be identified as disease risk factors. However, when viewed collectively, the importance of 1 of these may be reduced if the 2 attributes are highly correlated. The third criterion for feature inclusion was high frequency in the data set. Some of the laboratory values, particularly serum fibrinogen, were so sparse in the data set that exclusion of the feature was preferable to the alternatives of sample reduction or interpolation. Two thresholds for feature sparsity were established in this project. Models were built upon

data sets that excluded features with more than 20% missing values. Feature engineering substantially benefits from guidance by clinical domain experts [42], and our feature and outcome engineering was clinically informed, particularly in the realm of the diagnostic criteria described earlier. The features included in the final XGBoost model, the top-performing strategy, are available as supplementary materials to this manuscript. XGBoost is a DT-based ensemble modeling method. It can effectively capture the nonlinear relationship between predictors and the outcome by combining many weaker models to create a strong model. “Weak” and “strong” here refer to how correlated the models are to the outcome. The algorithm added models sequentially, and the next model corrected the error from the previous model. Through this iterative process, the data can be eventually accurately predicted by the model.

Usage Data and Generalizability

The application of usage data to this effort is both a weakness and a strength of this project. These data do not contain the richness of a complete medical record. It is therefore impossible to establish the criteria under which the clinicians made the diagnoses recorded—hence our outcome engineering maneuvers to establish stricter criteria (eg, using CPT codes for cataract surgery to identify patients with clinically significant cataract). At the same time, models built upon these sorts of data are more generalizable and available than models built upon more specific and perhaps more idiosyncratic data sources. These are precisely the sorts of data available to PCPs, making these models more easily deployable than models built upon a specific medical record system. Indeed, the availability of these data is illustrated by our being able to investigate a base of more than 80 million patients from disparate health care systems.

Definitions of the parameters used in these models is a topic worth addressing. The parameters ingested by the models that are used to make predictions include pathologies and demographics that would ordinarily require a clear and consistent definition. These parameters include macular degeneration whose definition should be established a priori to demographic terms, such as gender and sex, that not only require definition but also incorporate the idea of nonbinary values.

It is the nature of large electronic medical record studies that such definitions are impossible to impose externally and that the interpretations of gender, hypertension, diabetes, and glaucoma are likely to vary among the practitioners and patients who themselves may be the source of the data of these values in the data set. Our use of a database of 80 million patients provides a large degree of protection from selection bias. However, because these clinical definitions are intrinsic to the data set itself, a great deal of caution must be exercised when attempting to draw inferences about pathogenesis simply by evaluating the most correlative features of the model. However, the limitation of the model to revealing the disease process makes the model no less valuable in its ability to predict which patients are at the highest risk for unrecognized eye disease.

Hierarchical Relationships

It should be noted that the clinical features identified as relevant by each of the pathology models should be viewed as correlative

but not necessarily causative. It is better to think of the collection of clinical values as a patient's *clinical milieu* rather than as a collection of individual risk factors. Although it is difficult to imagine that college education is itself a risk factor for pathology, its correlation and importance to a given model should not be discounted, since it does contribute to the model's predictiveness of the presence of pathology. All of this is not to say that causation may not exist in the relation between some of these features and the pathologies modeled. Highly multidimensional clinical AI studies like this one may identify previously unrecognized factors that directly influence pathogenesis. However, causative connection cannot be established by this sort of study and would require a more traditional experimental approach. Although the J48 DT models did not perform as well as the GLM or XGBoost strategies, they are informative in that they describe hierarchical relationships among clinical features. As an example, the J48 model for glaucoma identifies race, systemic steroids, and antidiabetic medication use as important clinical features. However, the model dictates the order in which these factors should be considered, assessing race only after it is established whether the patient takes antidiabetic medications and assessing systemic steroid use only after these first 2 attributes have been determined. Such a hierarchical relationship among clinical features and demographic characteristics would be enormously difficult to establish in traditional reduced-dimensional scientific queries. This gestalt approach to multidimensional clinical context is one of the strengths of AI.

Decision Support

Ophthalmology is well suited for AI, given the rich visual information and data available; complex ophthalmological systems are better understood and eye care enhanced through sophisticated analysis and prediction. Integrating AI into clinical practice may facilitate better patient outcomes, given the complexity of disease diagnosis, treatment selection, and clinical testing. Ophthalmological clinical decision support systems that aid in diagnosis could improve the accuracy and efficiency of decision-making processes in ophthalmology, ultimately leading to improved patient access, outcomes, and potentially costs [43].

These models predict the presence of extant pathology. They would be of value in the identification of populations in which these pathologies are substantially more prevalent than in the general population. The models should not be used to make a diagnosis for an individual patient but rather to identify patients at risk of having undetected AMD, cataract, DR, glaucoma, or OSD. Further, these models are built upon clinical data in which an ophthalmic pathology is or is not present. That is to say, the models presented here are not constructed to predict the development of future pathology. It may or may not be the case that a particular clinical context, as defined by the multidimensional features incorporated into the models, may predict the development of future disease, but that is not appropriate way to use the models presented. These models predict the presence of ophthalmic pathology based upon non-ophthalmic data and would be best used for triage and referrals from non-ophthalmologists to eye care specialists. The research is designed to raise awareness about the variables associated with referral to heighten PCPs' vigilance to the

clinical and demographic characteristics that may need further reflection and attention.

Real-World Application Prospects of Ophthalmological AI Models

Advances in computing power combined with disruptions in health care resulting from unprecedented circumstances of the COVID-19 pandemic have prompted the worldwide exploration of AI-based systems in several medical subfields, including ophthalmology [44]. Ophthalmology has been at the forefront of AI research, in particular ML and DL approaches, because of the ubiquitous availability of noninvasive, rapid, and relatively inexpensive ophthalmic imaging [45]. Ophthalmic AI systems are advantageous in that they decrease the amount of time required to interpret image data, enable ophthalmologists to gain a greater understanding of disease progression, and assist with early-stage diagnosis, staging, and prognosis [46].

Numerous factors will determine the successful adoption of AI technologies into clinical practice. AI innovations that help clinicians manage the complexity (rather than add yet another layer of complexity) associated with effective ophthalmological care will likely be better received. In addition, the ability for critical appraisals by optometrists and ophthalmologists will be key to validating the theoretic models. AI models can be difficult to interpret and explain, which can make it difficult for stakeholders to understand how decisions are made [47]. It is important that the AI models be transparent and explainable in order to gain and maintain the trust of health care professionals, patients, and other decision makers. Providers of AI technologies and educators also need to ensure that training needs are adequately assessed and value to patient outcomes demonstrated if the promise of AI in ophthalmological care is to be realized.

AI has the potential to provide invaluable insights across multiple domains of ophthalmology. By leveraging ML algorithms, AI can process and analyze vast amounts of information, including physiological data, EHRs, 3D images, radiology images, histologic evaluation, genomic sequencing, and administrative and billing data. One advantage that could be realized by the algorithms discussed herein is that they use commonly collected data contained within an EHR system to identify eye disease risk. This means that the algorithms could be deployed in the background of an EHR to enable inference of an entire PCP's or practice's patient population. The results of this inference could appear as a flag in a patient chart, alerting the PCP for a given patient as to the need to refer to an eye care professional for further evaluation. The approach of deploying these algorithms within the EHR would also enable further validation and assessment of algorithm generalizability prior to clearing the algorithm for regular use by PCPs. Additional validation steps such as this would help identify any local biases for a given patient population and enable monitoring performance for algorithmic drift.

Data infrastructure is an important influencer for the adoption of AI innovations. AI requires a continuous supply of high-quality data. Data quality issues may entail accuracy, completeness, consistency, timeliness, integrity, relevance, data collection, preprocessing, management, data governance, and data labeling [47]. Storage challenges, processing challenges,

data management challenges, data heterogeneity, data privacy and security, bias and representativeness, and data access are also data quality considerations [47]. An appropriate data infrastructure, including its maintenance and evolution over time, is a prerequisite for successful AI applications.

Management of eye health necessitates a multidisciplinary team with a dynamic flow of information between treating doctors [48]. Holley and Lee's [4] qualitative research found that PCPs had poor communication with eye care providers and the PCPs desire changes in the current referral-to-eye-care system. Better communication between PCPs and eye care professionals, further implementation of EHRs, and increasing eye screening in primary care clinics were common themes. Moudgil et al [48] found that 80% of the physicians communicated with ophthalmologists sometimes, whereas only 10% ensured communication at all times. The information sought by the treating physicians from the ophthalmologists regarding their referral for ocular findings included severity, the grading of DR, other ocular changes, need for intervention, and the frequency of screening and follow-up based on changes observed.

Finally, ethical considerations call for AI systems to adhere to the principles of fairness and nondiscrimination [49,50]. Advances in modern medicine are sometimes stymied by the inability to translate evidence-based care to all patients [51]. Transparency of AI models is essential to be able to evaluate and ensure their relevance for diverse populations and the ability to translate the innovations to all settings of care.

Limitations

Several limitations are inherent in the use of aggregated clinical data. Longitudinal data on patients are limited, and this, by extension, limits projects such as ours in their ability to predict the development of future pathology. Although the data set does derive information from EHRs, including Epic, Cerner, GE, and McKesson, the actual physicians' notes are not available for analysis. Aggregated data also disproportionately represent

hospital encounters and underrepresent outpatient visits [52]. Attempts to mitigate some of these deficiencies in the feature and outcome engineering methods are described before. A certain degree of circumspection should be exercised when applying this model more broadly to other databases that may have used different NLP protocols.

A challenge with deploying these models in their current form is that the richness of data (ie, number of parameters) to be input into the models must be balanced against the labor the clinician must expend entering them. The authors sought to reduce feature input without substantially affecting model predictive performance. The goal is to develop tools that will aid clinicians and reduce the number of undiagnosed serious ophthalmic conditions. Empirically based analyses such those presented here are exploratory and intended to generate insights worthy of subsequent investigation with different study designs and methods that are better suited for causal inference.

It is important to note that data quality and representativeness are a potential issue for ML model training from EHRs and other clinical databases. EHR data can be incomplete, inconsistent, or erroneous, given the nature of the data collection and documentation. EHR data can also be biased toward populations with better access to health care. Some of these issues (eg, access) are inherent to our health care system in general and are not specific to EHR data. Regardless of the source of the issue, it is important to note that models trained and tested on EHR data may not be generalizable to the larger population.

Conclusion

In summary, this research demonstrates real patient triage potential by deploying AI strategies directly to PCP EHRs. In addition, based on the original data pool (more than 80 million patients), the final study population size (1,486,078 patients, 50% of whom were controls) and the 9 subpathologies using 5 different analytical modeling approaches, the authors believe this study to be one of the largest AI projects in ophthalmology.

Acknowledgments

This study was funded by Johnson & Johnson Vision, Inc. The sponsor participated in the design of the study, conducting the study, data collection, data management, data analysis, interpretation of the data, and preparation, review, and approval of the manuscript.

Conflicts of Interest

JAY is a consultant for Johnson & Johnson Vision, Inc. CWC, CWS, CEH and CAB are employees of Johnson & Johnson. SVM was a contractor with Johnson & Johnson at the time of the study.

Multimedia Appendix 1

Patient inclusion and exclusion criteria and attrition.

[[DOCX File, 16 KB - ai_v3i1e48295_app1.docx](#)]

References

1. Saydah SH, Gerzoff RB, Saaddine JB, Zhang X, Cotch MF. Eye care among US adults at high risk for vision loss in the United States in 2002 and 2017. *JAMA Ophthalmol* 2020 May 01;138(5):479-489 [FREE Full text] [doi: [10.1001/jamaophthalmol.2020.0273](https://doi.org/10.1001/jamaophthalmol.2020.0273)] [Medline: [32163124](https://pubmed.ncbi.nlm.nih.gov/32163124/)]

2. Varma R, Vajaranant TS, Burkemper B, Wu S, Torres M, Hsu C, et al. Visual impairment and blindness in adults in the United States: demographic and geographic variations from 2015 to 2050. *JAMA Ophthalmol* 2016 Jul 01;134(7):802-809 [FREE Full text] [doi: [10.1001/jamaophthalmol.2016.1284](https://doi.org/10.1001/jamaophthalmol.2016.1284)] [Medline: [27197072](https://pubmed.ncbi.nlm.nih.gov/27197072/)]
3. Whitson HE, Cousins SW, Burchett BM, Hybels CF, Pieper CF, Cohen HJ. The combined effect of visual impairment and cognitive impairment on disability in older people. *J Am Geriatr Soc* 2007 Jun 25;55(6):885-891. [doi: [10.1111/j.1532-5415.2007.01093.x](https://doi.org/10.1111/j.1532-5415.2007.01093.x)] [Medline: [17537089](https://pubmed.ncbi.nlm.nih.gov/17537089/)]
4. Holley CD, Lee PP. Primary care provider views of the current referral-to-eye-care process: focus group results. *Invest Ophthalmol Vis Sci* 2010 Apr 01;51(4):1866-1872. [doi: [10.1167/iovs.09-4512](https://doi.org/10.1167/iovs.09-4512)] [Medline: [19875660](https://pubmed.ncbi.nlm.nih.gov/19875660/)]
5. Rhodes L, Huisingh C, McGwin G, Mennemeyer S, Bregantini M, Patel N, et al. Eye Care Quality and Accessibility Improvement in the Community (EQUALITY): impact of an eye health education program on patient knowledge about glaucoma and attitudes about eye care. *Patient Relat Outcome Meas* 2016 May;7:37-48. [doi: [10.2147/prom.s98686](https://doi.org/10.2147/prom.s98686)]
6. Paz SH, Varma R, Klein R, Wu J, Azen SP, Los Angeles Latino Eye Study Group. Noncompliance with vision care guidelines in Latinos with type 2 diabetes mellitus: the Los Angeles Latino Eye Study. *Ophthalmology* 2006 Aug;113(8):1372-1377. [doi: [10.1016/j.ophtha.2006.04.018](https://doi.org/10.1016/j.ophtha.2006.04.018)] [Medline: [16769120](https://pubmed.ncbi.nlm.nih.gov/16769120/)]
7. McMonnies CW. Glaucoma history and risk factors. *J Optom* 2017 Apr;10(2):71-78 [FREE Full text] [doi: [10.1016/j.optom.2016.02.003](https://doi.org/10.1016/j.optom.2016.02.003)] [Medline: [27025415](https://pubmed.ncbi.nlm.nih.gov/27025415/)]
8. Mitchell P, Lee AJ, Wang JJ, Rochtchina E. Intraocular pressure over the clinical range of blood pressure: Blue Mountains Eye Study findings. *Am J Ophthalmol* 2005 Jul;140(1):131-132. [doi: [10.1016/j.ajo.2004.12.088](https://doi.org/10.1016/j.ajo.2004.12.088)] [Medline: [16038656](https://pubmed.ncbi.nlm.nih.gov/16038656/)]
9. Zhou M, Wang W, Huang W, Zhang X. Diabetes mellitus as a risk factor for open-angle glaucoma: a systematic review and meta-analysis. *PLoS One* 2014 Aug 19;9(8):e102972 [FREE Full text] [doi: [10.1371/journal.pone.0102972](https://doi.org/10.1371/journal.pone.0102972)] [Medline: [25137059](https://pubmed.ncbi.nlm.nih.gov/25137059/)]
10. Pérez-de-Arcelus M, Toledo E, Martínez-González M, Martín-Calvo N, Fernández-Montero A, Moreno-Montañés J. Smoking and incidence of glaucoma: the SUN Cohort. *Medicine (Baltimore)* 2017;96:e5761. [doi: [10.1097/md.00000000000005761](https://doi.org/10.1097/md.00000000000005761)]
11. Gordon MO, Beiser JA, Brandt JD, Heuer DK, Higginbotham EJ, Johnson CA, et al. The Ocular Hypertension Treatment Study: baseline factors that predict the onset of primary open-angle glaucoma. *Arch Ophthalmol* 2002 Jun 01;120(6):714-720; discussion 729-730. [doi: [10.1001/archophth.120.6.714](https://doi.org/10.1001/archophth.120.6.714)] [Medline: [12049575](https://pubmed.ncbi.nlm.nih.gov/12049575/)]
12. Glynn RJ, Rosner B, Christen WG. Evaluation of risk factors for cataract types in a competing risks framework. *Ophthalmic Epidemiol* 2009 Jul 08;16(2):98-106 [FREE Full text] [doi: [10.1080/09286580902737532](https://doi.org/10.1080/09286580902737532)] [Medline: [19353398](https://pubmed.ncbi.nlm.nih.gov/19353398/)]
13. Glynn RJ, Christen WG, Manson JE, Bernheimer J, Hennekens CH. Body mass index. An independent predictor of cataract. *Arch Ophthalmol* 1995 Sep 01;113(9):1131-1137. [doi: [10.1001/archophth.1995.01100090057023](https://doi.org/10.1001/archophth.1995.01100090057023)] [Medline: [7661746](https://pubmed.ncbi.nlm.nih.gov/7661746/)]
14. Zhang G, Chen H, Chen W, Zhang M. Prevalence and risk factors for diabetic retinopathy in China: a multi-hospital-based cross-sectional study. *Br J Ophthalmol* 2017 Dec 30;101(12):1591-1595 [FREE Full text] [doi: [10.1136/bjophthalmol-2017-310316](https://doi.org/10.1136/bjophthalmol-2017-310316)] [Medline: [28855195](https://pubmed.ncbi.nlm.nih.gov/28855195/)]
15. Chakravarthy U, Wong TY, Fletcher A, Piau E, Evans C, Zlateva G, et al. Clinical risk factors for age-related macular degeneration: a systematic review and meta-analysis. *BMC Ophthalmol* 2010 Dec 13;10(1):31 [FREE Full text] [doi: [10.1186/1471-2415-10-31](https://doi.org/10.1186/1471-2415-10-31)] [Medline: [21144031](https://pubmed.ncbi.nlm.nih.gov/21144031/)]
16. Yang W, Yang Y, Cao J. Risk factors for dry eye syndrome: a retrospective case-control study. *Optom Vis Sci* 2015;92:e199-e205. [doi: [10.1097/OPX.0000000000000541](https://doi.org/10.1097/OPX.0000000000000541)]
17. Krittanawong C, Zhang H, Wang Z, Aydar M, Kitai T. Artificial intelligence in precision cardiovascular medicine. *J Am Coll Cardiol* 2017 May 30;69(21):2657-2664 [FREE Full text] [doi: [10.1016/j.jacc.2017.03.571](https://doi.org/10.1016/j.jacc.2017.03.571)] [Medline: [28545640](https://pubmed.ncbi.nlm.nih.gov/28545640/)]
18. Nensa F, Demircioglu A, Rischpler C. Artificial intelligence in nuclear medicine. *J Nucl Med* 2019 Sep 03;60(Suppl 2):29S-37S [FREE Full text] [doi: [10.2967/jnumed.118.220590](https://doi.org/10.2967/jnumed.118.220590)] [Medline: [31481587](https://pubmed.ncbi.nlm.nih.gov/31481587/)]
19. Stewart J, Sprivilis P, Dwivedi G. Artificial intelligence and machine learning in emergency medicine. *Emerg Med Australas* 2018 Dec 16;30(6):870-874. [doi: [10.1111/1742-6723.13145](https://doi.org/10.1111/1742-6723.13145)] [Medline: [30014578](https://pubmed.ncbi.nlm.nih.gov/30014578/)]
20. Schork N. Artificial intelligence and personalized medicine. *Cancer Treat Res* 2019;178:265-283 [FREE Full text] [doi: [10.1007/978-3-030-16391-4_11](https://doi.org/10.1007/978-3-030-16391-4_11)] [Medline: [31209850](https://pubmed.ncbi.nlm.nih.gov/31209850/)]
21. Cheng H, Kane JX, Liu L, Li J, Cheng B, Wu M. Refractive predictability using the IOLMaster 700 and artificial intelligence-based IOL power formulas compared to standard formulas. *J Refract Surg* 2020 Jul;36(7):466-472. [doi: [10.3928/1081597x-20200514-02](https://doi.org/10.3928/1081597x-20200514-02)]
22. Carmona González D, Palomino Bautista C. Accuracy of a new intraocular lens power calculation method based on artificial intelligence. *Eye* 2020 Apr 28;35(2):517-522 [FREE Full text] [doi: [10.1038/s41433-020-0883-3](https://doi.org/10.1038/s41433-020-0883-3)] [Medline: [32346109](https://pubmed.ncbi.nlm.nih.gov/32346109/)]
23. Kane J, Van Heerden A, Atik A, Petsoglou C. Accuracy of 3 new methods for intraocular lens power selection. *J Cataract Refract Surg* 2017 Mar;43(3):333-339. [doi: [10.1016/j.jcrs.2016.12.021](https://doi.org/10.1016/j.jcrs.2016.12.021)] [Medline: [28410714](https://pubmed.ncbi.nlm.nih.gov/28410714/)]
24. Devalla SK, Liang Z, Pham TH, Boote C, Strouthidis NG, Thiery AH, et al. Glaucoma management in the era of artificial intelligence. *Br J Ophthalmol* 2020 Mar 22;104(3):301-311. [doi: [10.1136/bjophthalmol-2019-315016](https://doi.org/10.1136/bjophthalmol-2019-315016)] [Medline: [31640973](https://pubmed.ncbi.nlm.nih.gov/31640973/)]
25. Song Y, Ishikawa H, Wu M, Liu Y, Lucy KA, Lavinsky F, et al. Clinical prediction performance of glaucoma progression using a 2-dimensional continuous-time hidden Markov model with structural and functional measurements. *Ophthalmology* 2018 Sep;125(9):1354-1361 [FREE Full text] [doi: [10.1016/j.ophtha.2018.02.010](https://doi.org/10.1016/j.ophtha.2018.02.010)] [Medline: [29571832](https://pubmed.ncbi.nlm.nih.gov/29571832/)]

26. Gulshan V, Peng L, Coram M, Stumpe MC, Wu D, Narayanaswamy A, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA* 2016 Dec 13;316(22):2402-2410. [doi: [10.1001/jama.2016.17216](https://doi.org/10.1001/jama.2016.17216)] [Medline: [27898976](https://pubmed.ncbi.nlm.nih.gov/27898976/)]
27. Wan Q, Tang J. Exploration of potential key pathways and genes in multiple ocular cancers through bioinformatics analysis. *Graefes Arch Clin Exp Ophthalmol* 2019 Oct 15;257(10):2329-2341. [doi: [10.1007/s00417-019-04410-2](https://doi.org/10.1007/s00417-019-04410-2)] [Medline: [31309275](https://pubmed.ncbi.nlm.nih.gov/31309275/)]
28. Cowie MR, Blomster JI, Curtis LH, Duclaux S, Ford I, Fritz F, et al. Electronic health records to facilitate clinical research. *Clin Res Cardiol* 2017 Jan 24;106(1):1-9 [FREE Full text] [doi: [10.1007/s00392-016-1025-6](https://doi.org/10.1007/s00392-016-1025-6)] [Medline: [27557678](https://pubmed.ncbi.nlm.nih.gov/27557678/)]
29. Birtwhistle R, Williamson T. Primary care electronic medical records: a new data source for research in Canada. *CMAJ* 2015 Mar 03;187(4):239-240 [FREE Full text] [doi: [10.1503/cmaj.140473](https://doi.org/10.1503/cmaj.140473)] [Medline: [25421989](https://pubmed.ncbi.nlm.nih.gov/25421989/)]
30. Schuemie MJ, Cepeda MS, Suchard MA, Yang J, Tian Y, Schuler A, et al. How confident are we about observational findings in healthcare: a benchmark study. *Harv Data Sci Rev* 2020 Jan 31;2(1):1771-1781 [FREE Full text] [doi: [10.1162/99608f92.147cc28e](https://doi.org/10.1162/99608f92.147cc28e)] [Medline: [33367288](https://pubmed.ncbi.nlm.nih.gov/33367288/)]
31. Zheng B, Agresti A. Summarizing the predictive power of a generalized linear model. *Stat Med* 2000 Jul 15;19(13):1771-1781. [doi: [10.1002/1097-0258\(20000715\)19:13<1771::aid-sim485>3.0.co;2-p](https://doi.org/10.1002/1097-0258(20000715)19:13<1771::aid-sim485>3.0.co;2-p)] [Medline: [10861777](https://pubmed.ncbi.nlm.nih.gov/10861777/)]
32. Lee S, Lee H, Abbeel P, Ng A. EfficientL1 regularized logistic regression. 2006 Presented at: AAI'06: 21st National Conference on Artificial intelligence; July 16-20, 2006; Boston, MA.
33. Speiser JL, Miller ME, Tooze J, Ip E. A comparison of random forest variable selection methods for classification prediction modeling. *Expert Syst Appl* 2019 Nov 15;134:93-101 [FREE Full text] [doi: [10.1016/j.eswa.2019.05.028](https://doi.org/10.1016/j.eswa.2019.05.028)] [Medline: [32968335](https://pubmed.ncbi.nlm.nih.gov/32968335/)]
34. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. 2016 Presented at: KDD '16: 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; August 13-17, 2016; San Francisco, CA. [doi: [10.1145/2939672.2939785](https://doi.org/10.1145/2939672.2939785)]
35. Kaur G, Chhabra A. Improved J48 classification algorithm for the prediction of diabetes. *Int J Comput Appl* 2014 Jul 18;98(22):13-17. [doi: [10.5120/17314-7433](https://doi.org/10.5120/17314-7433)]
36. Hogue C, Gaylor D, Schulz K. Estimators of relative risk for case-control studies. *Am J Epidemiol* 1983 Sep;118(3):396-407. [doi: [10.1093/oxfordjournals.aje.a113646](https://doi.org/10.1093/oxfordjournals.aje.a113646)] [Medline: [6613982](https://pubmed.ncbi.nlm.nih.gov/6613982/)]
37. Jin K, Ye J. Artificial intelligence and deep learning in ophthalmology: current status and future perspectives. *Adv Ophthalmol Pract Res* 2022 Nov;2(3):100078 [FREE Full text] [doi: [10.1016/j.aopr.2022.100078](https://doi.org/10.1016/j.aopr.2022.100078)] [Medline: [37846285](https://pubmed.ncbi.nlm.nih.gov/37846285/)]
38. Palestine AG, Merrill PT, Saleem SM, Jabs DA, Thorne JE. Assessing the precision of ICD-10 codes for uveitis in 2 electronic health record systems. *JAMA Ophthalmol* 2018 Oct 01;136(10):1186-1190 [FREE Full text] [doi: [10.1001/jamaophthalmol.2018.3001](https://doi.org/10.1001/jamaophthalmol.2018.3001)] [Medline: [30054618](https://pubmed.ncbi.nlm.nih.gov/30054618/)]
39. Hreha KP, Fisher SR, Reistetter TA, Ottenbacher K, Haas A, Li C, et al. Use of the ICD-10 vision codes to study ocular conditions in Medicare beneficiaries with stroke. *BMC Health Serv Res* 2020 Jul 08;20(1):628 [FREE Full text] [doi: [10.1186/s12913-020-05484-z](https://doi.org/10.1186/s12913-020-05484-z)] [Medline: [32641050](https://pubmed.ncbi.nlm.nih.gov/32641050/)]
40. Hellman JB, Lim M, Leung K, Blount C, Yiu G. The impact of conversion to International Classification of Diseases, 10th revision (ICD-10) on an academic ophthalmology practice. *Clin Ophthalmol* 2018 May;12:949-956. [doi: [10.2147/ophth.s161742](https://doi.org/10.2147/ophth.s161742)]
41. Syarif I, Prugel-Bennett A, Wills G. SVM parameter optimization using grid search and genetic algorithm to improve classification performance. *TELKOMNIKA* 2016 Dec 01;14(4):1502. [doi: [10.12928/telkomnika.v14i4.3956](https://doi.org/10.12928/telkomnika.v14i4.3956)]
42. Roe KD, Jawa V, Zhang X, Chute CG, Epstein JA, Matelsky J, et al. Feature engineering with clinical expert knowledge: a case study assessment of machine learning model complexity and performance. *PLoS One* 2020 Apr 23;15(4):e0231300 [FREE Full text] [doi: [10.1371/journal.pone.0231300](https://doi.org/10.1371/journal.pone.0231300)] [Medline: [32324754](https://pubmed.ncbi.nlm.nih.gov/32324754/)]
43. Comito C, Falcone D, Forestiero A. AI-driven clinical decision support: enhancing disease diagnosis exploiting patients similarity. *IEEE Access* 2022;10:6878-6888. [doi: [10.1109/access.2022.3142100](https://doi.org/10.1109/access.2022.3142100)]
44. Ahuja AS, Reddy VP, Marques O. Artificial intelligence and COVID-19: a multidisciplinary approach. *Integr Med Res* 2020 Sep;9(3):100434 [FREE Full text] [doi: [10.1016/j.imr.2020.100434](https://doi.org/10.1016/j.imr.2020.100434)] [Medline: [32632356](https://pubmed.ncbi.nlm.nih.gov/32632356/)]
45. Lee CS, Brandt JD, Lee AY. Big data and artificial intelligence in ophthalmology: where are we now? *Ophthalmol Sci* 2021 Jun;1(2):100036 [FREE Full text] [doi: [10.1016/j.xops.2021.100036](https://doi.org/10.1016/j.xops.2021.100036)] [Medline: [36249294](https://pubmed.ncbi.nlm.nih.gov/36249294/)]
46. Ahuja AS, Wagner IV, Dorairaj S, Checo L, Hulzen RT. Artificial intelligence in ophthalmology: a multidisciplinary approach. *Integr Med Res* 2022 Dec;11(4):100888 [FREE Full text] [doi: [10.1016/j.imr.2022.100888](https://doi.org/10.1016/j.imr.2022.100888)] [Medline: [36212633](https://pubmed.ncbi.nlm.nih.gov/36212633/)]
47. Aldoseri A, Al-Khalifa KN, Hamouda AM. Re-thinking data strategy and integration for artificial intelligence: concepts, opportunities, and challenges. *Appl Sci* 2023 Jun 13;13(12):7082. [doi: [10.3390/app13127082](https://doi.org/10.3390/app13127082)]
48. Moudgil T, Bains B, Bandhu S, Kanda N. Preferred practice pattern of physicians regarding diabetic retinopathy in diabetes mellitus patients. *Indian J Ophthalmol* 2021;69(11):3139-3143. [doi: [10.4103/ijo.ijo_1339_21](https://doi.org/10.4103/ijo.ijo_1339_21)]
49. Jobin A, Ienca M, Vayena E. The global landscape of AI ethics guidelines. *Nat Mach Intell* 2019 Sep 02;1(9):389-399. [doi: [10.1038/s42256-019-0088-2](https://doi.org/10.1038/s42256-019-0088-2)]
50. Weidener L, Fischer M. Role of ethics in developing AI-based applications in medicine: insights from expert interviews and discussion of implications. *JMIR AI* 2024 Jan 12;3:e51204. [doi: [10.2196/51204](https://doi.org/10.2196/51204)]

51. Khera R, Butte AJ, Berkwits M, Hswen Y, Flanagan A, Park H, et al. AI in medicine—JAMA’s focus on clinical outcomes, patient-centered care, quality, and equity. *JAMA* 2023 Sep 05;330(9):818-820. [doi: [10.1001/jama.2023.15481](https://doi.org/10.1001/jama.2023.15481)] [Medline: [37566406](https://pubmed.ncbi.nlm.nih.gov/37566406/)]
52. Rolnick J. Aggregate health data in the United States: steps toward a public good. *Health Informat J* 2013 Jun 27;19(2):137-151 [FREE Full text] [doi: [10.1177/1460458212462077](https://doi.org/10.1177/1460458212462077)] [Medline: [23715213](https://pubmed.ncbi.nlm.nih.gov/23715213/)]

Abbreviations

AI: artificial intelligence
AMD: age-related macular degeneration
AUC: area under the curve
DEMGD: dry eye and meibomian gland dysfunction
DL: deep learning
DR: diabetic retinopathy
DT: decision tree
EHR: electronic health record
GLM: generalized linear model
ICD-10: International Classification of Diseases, 10th Revision
LR: logistic regression without regularization
L1-LR: L1-regularized logistic regression
ML: machine learning
NLP: natural language processing
NPDR: nonproliferative diabetic retinopathy
OR: odds ratio
OSD: ocular surface disease
PCP: primary care practitioner
PDR: proliferative diabetic retinopathy
RF: random forest
XGBoost: Extreme Gradient Boosting

Edited by K El Emam, B Malin; submitted 18.04.23; peer-reviewed by E Shaveet, Y Chu, N Soley; comments to author 19.05.23; revised version received 11.07.23; accepted 10.02.24; published 12.03.24.

Please cite as:

Young JA, Chang CW, Scales CW, Menon SV, Holy CE, Blackie CA

Machine Learning Methods Using Artificial Intelligence Deployed on Electronic Health Record Data for Identification and Referral of At-Risk Patients From Primary Care Physicians to Eye Care Specialists: Retrospective, Case-Controlled Study
JMIR AI 2024;3:e48295

URL: <https://ai.jmir.org/2024/1/e48295>

doi: [10.2196/48295](https://doi.org/10.2196/48295)

PMID: [38875582](https://pubmed.ncbi.nlm.nih.gov/38875582/)

©Joshua A Young, Chin-Wen Chang, Charles W Scales, Saurabh V Menon, Chantal E Holy, Caroline Adrienne Blackie. Originally published in *JMIR AI* (<https://ai.jmir.org>), 12.03.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in *JMIR AI*, is properly cited. The complete bibliographic information, a link to the original publication on <https://www.ai.jmir.org/>, as well as this copyright and license information must be included.

Original Paper

Risk Perception, Acceptance, and Trust of Using AI in Gastroenterology Practice in the Asia-Pacific Region: Web-Based Survey Study

Wilson WB Goh^{1,2,3}, BSc, MSc, PhD; Kendrick YA Chia^{1,2,3}, BAcc, BBM, MSc; Max FK Cheung¹, BSc; Kalya M Kee^{1,4}, BA, MSc; May O Lwin⁴, BA, MBA, PhD; Peter J Schulz^{1,4}, BA, MA, PhD; Minhu Chen⁵, MBBS, PhD; Kaichun Wu⁶, MD, PhD; Simon SM Ng⁷, MBChB, MD; Rashid Lui⁸, MBChB; Tiing Leong Ang⁹, MBBS, MRCPUK; Khay Guan Yeoh^{10,11}, MBBS, MMed; Han-mo Chiu^{12,13}, MD, PhD; Deng-chyang Wu¹⁴, MD, PhD; Joseph JY Sung¹, MD, PhD

¹Lee Kong Chian School of Medicine, Nanyang Technological University, Singapore, Singapore, Singapore

²School of Biological Sciences, Nanyang Technological University, Singapore, Singapore

³Center for Biomedical Informatics, Nanyang Technological University, Singapore, Singapore

⁴Wee Kim Wee School of Communication and Information, Nanyang Technological University, Singapore, Singapore

⁵The First Affiliated Hospital, Sun Yat-sen University, Guangzhou, China

⁶Xijing Hospital, Fourth Military Medical University, Xi'an, China

⁷Department of Surgery, Faculty of Medicine, The Chinese University of Hong Kong, Hong Kong, China (Hong Kong)

⁸Prince of Wales Hospital, Hospital Authority, Hong Kong, China (Hong Kong)

⁹Department of Gastroenterology and Hepatology, Changi General Hospital, SingHealth, Singapore, Singapore

¹⁰Department of Gastroenterology and Hepatology, National University Hospital, National University Health System, Singapore, Singapore

¹¹Department of Medicine, Yong Loo Lin School of Medicine, National University of Singapore, Singapore, Singapore

¹²Department of Internal Medicine, National Taiwan University Hospital, Taiwan, China

¹³Department of Internal Medicine, College of Medicine, National Taiwan University, Taiwan, China

¹⁴Kaohsiung Medical University, Taiwan, China

Corresponding Author:

Wilson WB Goh, BSc, MSc, PhD

Lee Kong Chian School of Medicine

Nanyang Technological University

Singapore

Experimental Medicine Building, 59 Nanyang Drive

Singapore, 636921

Singapore

Phone: 65 67911744

Email: wilsongoh@ntu.edu.sg

Abstract

Background: The use of artificial intelligence (AI) can revolutionize health care, but this raises risk concerns. It is therefore crucial to understand how clinicians trust and accept AI technology. Gastroenterology, by its nature of being an image-based and intervention-heavy specialty, is an area where AI-assisted diagnosis and management can be applied extensively.

Objective: This study aimed to study how gastroenterologists or gastrointestinal surgeons accept and trust the use of AI in computer-aided detection (CADE), computer-aided characterization (CADx), and computer-aided intervention (CADi) of colorectal polyps in colonoscopy.

Methods: We conducted a web-based questionnaire from November 2022 to January 2023, involving 5 countries or areas in the Asia-Pacific region. The questionnaire included variables such as background and demography of users; intention to use AI, perceived risk; acceptance; and trust in AI-assisted detection, characterization, and intervention. We presented participants with 3 AI scenarios related to colonoscopy and the management of colorectal polyps. These scenarios reflect existing AI applications in colonoscopy, namely the detection of polyps (CADE), characterization of polyps (CADx), and AI-assisted polypectomy (CADi).

Results: In total, 165 gastroenterologists and gastrointestinal surgeons responded to a web-based survey using the structured questionnaire designed by experts in medical communications. Participants had a mean age of 44 (SD 9.65) years, were mostly male (n=116, 70.3%), and mostly worked in publicly funded hospitals (n=110, 66.67%). Participants reported relatively high exposure to AI, with 111 (67.27%) reporting having used AI for clinical diagnosis or treatment of digestive diseases. Gastroenterologists are highly interested to use AI in diagnosis but show different levels of reservations in risk prediction and acceptance of AI. Most participants (n=112, 72.72%) also expressed interest to use AI in their future practice. CADE was accepted by 83.03% (n=137) of respondents, CADx was accepted by 78.79% (n=130), and CADi was accepted by 72.12% (n=119). CADE and CADx were trusted by 85.45% (n=141) of respondents and CADi was trusted by 72.12% (n=119). There were no application-specific differences in risk perceptions, but more experienced clinicians gave lesser risk ratings.

Conclusions: Gastroenterologists reported overall high acceptance and trust levels of using AI-assisted colonoscopy in the management of colorectal polyps. However, this level of trust depends on the application scenario. Moreover, the relationship among risk perception, acceptance, and trust in using AI in gastroenterology practice is not straightforward.

(JMIR AI 2024;3:e50525) doi:[10.2196/50525](https://doi.org/10.2196/50525)

KEYWORDS

artificial intelligence; delivery of health care; gastroenterology; acceptance; trust; adoption; survey; surveys; questionnaire; questionnaires; detect; detection; colonoscopy; gastroenterologist; gastroenterologists; internal medicine; polyp; polyps; surgeon; surgeons; surgery; surgical; colorectal

Introduction

Artificial intelligence (AI) has made groundbreaking technological advancements in medical image interpretation [1]; diagnosis assistance; risk assessment for various conditions [2]; outcome prognostication [3]; and in certain areas, treatment suggestion [4] and partaking in surgical intervention [5].

Studies of AI trust and acceptance among clinicians are becoming increasingly important. This is because trust and acceptance of AI technology are seen as preconditions for clinical workflow integration [6]. Currently, trust has already been demonstrated by several studies as one of the main determinants in driving the adoption of AI in health care [7,8]. One study showed that within a general home-based health care setting—where AI is applied on the internet of things–based devices to monitor patients’ health—risk perception, acceptance, and trust are related concepts that govern the ultimate use of the developed technology [9]. A separate study [10] conducted on the use of an AI-based system in the application of a Blood Utilization Calculator showed that its trust and use were determined by perceived risk and expectancy (in our context, acceptance). It was demonstrated that high perceived risk reduced trust and subsequent use.

While the clinical evidence of accuracy in the diagnosis and prognosis of AI is accumulating, the level of trust and acceptance by clinicians requires more attention [6]. We identified that gastroenterology, by its very nature of having heavy usage of image-based diagnosis (eg, computed tomography, magnetic resonance imaging, endoscopy, and histology) and surgical or endoscopic intervention, will be one of the specialties that may readily use AI technologies in clinical management [11,12]. Yet, there is little research on AI risk perception, acceptance, and trust among gastroenterologists.

To our knowledge, most published research surveys trust in a more general manner. One such recent example is the survey on gastrointestinal (GI) health care in 2022, which covered clinicians’ perspectives in a general way [13]. However, such

surveys lack granularity. It is impossible to know under what circumstances do clinicians become less trusting or accepting or become more concerned about the deployments of AI.

Moreover, there is a lack of explicit modeling from collected data to relate patterns of risk perception, acceptance, and trust among practitioners. There are existing models [14,15] that explore parts of the interactions among these 3 factors. However, because these explorations cover only partial relationships and interactions, we feel that these may be inadequate for modeling real-world dynamics. Therefore, having more comprehensive models would allow for a better understanding of the various factors underpinning how clinicians come to trust, accept, and eventually use AI. This knowledge would help in formulating successful implementation of AI tools in real-world environments.

In this study, we aim to understand the trust and acceptance among gastroenterologists, with a specific focus on the Asia-3Pacific region. We hypothesize is that risk perception, acceptance, and trust will change according to the scenario (computer-aided detection [CADE], computer-aided characterization [CADx], or computer-aided intervention [CADi]), with different levels of invasiveness. A blueprint of a survey that examines contextual responses toward screening colonoscopy with polypectomy in clinical environments is provided. Using our collected data, we attempt to elucidate how risk perception, acceptance, and trust interactions can be modeled and studied. These contributions collectively enhance our understanding of complex factors influencing the integration of AI in medical practice.

Methods

Survey

We used a structured questionnaire (Multimedia Appendix 1) to conduct a survey in English by inviting gastroenterologists or GI surgeons from the Asia-Pacific region through open invitations to various medical associations. The questionnaire was based on the expectancy-value framework, major constructs

of the Theory of Planned Behaviour research framework [16], and the Technology Acceptance Model measures [17]. Items in the questionnaire for testing risk perception, acceptance, and trust were adapted from various other studies [18,19], with some including items from validated constructs in questionnaires. These questions are then adapted into scenarios covering detection (CADE), characterization (CADx), or intervention (CADi), with different levels of invasiveness characterization and intervention for colonoscopic detection and polypectomy (see [Textbox 1](#) for items used to evaluate these aspects).

Most items were rated on a 7-point Likert scale, where 7 denotes strong agreement. To assess risk perception, acceptance, and trust, we presented participants with 3 different AI applications related to colonoscopy and the management of colorectal polyps. These scenarios, reflecting existing AI applications in GI, involve the detection of polyps (CADE), characterization of the nature of polyps (CADx), and treatment procedures (CADi), respectively (see [Table 1](#) and [Textbox 1](#)). [Table 2](#) displays measurement items.

In this study, the three key elements for assessment are (1) risk perception, (2) acceptance, and (3) trust. Risk perception refers to an individual's subjective assessment or understanding of

the potential hazards, threats, or uncertainties associated with a particular situation or activity. It involves the process of evaluating and interpreting information about risk, considering factors such as the severity of potential consequences [20,21]. Acceptance is the mental and emotional state of acknowledging and accommodating a new concept or innovation into one's beliefs, behaviors, or practices. Trust is defined as belief or confidence in the reliability, credibility, and integrity of a person, system, or technology leading to usage or action [20,21]. Acceptance may precede trust in the adoption of new technologies, but trust plays a crucial role in establishing a strong foundation for sustained usage and effective integration of AI into medical practice. Risk perception, acceptance, and trust may interact with each other and other factors stemming from professional, technological, and personal sources. The conceptual framework presented in [Figure 1](#) illustrates the intricate interplay among sociodemographic variables, AI acceptance, trust, perceived risk, and outcomes [22]. Our study aims to contribute to this understanding not by testing individual relationships within this conceptual framework but by exploring how trust, risk, and acceptance are possibly interconnected in the context of AI-supported applications in gastroenterology.

Textbox 1. The 3 operationalized case scenarios of using artificial intelligence–assisted colonoscopy in the management of colorectal polyps.

Computer-aided detection

- Imagine you are attending an informal meeting of colleagues. Your colleagues are not experts in artificial intelligence and have about the same amount of understanding as you do. The conversation turns to innovation in medicine, especially machine learning algorithms and their potential to assist in the interpretation of medical imagery in the early detection of colon cancer. One of the colleagues speaks about a patient who underwent a colonoscopy which was assisted by a machine learning algorithm. When the algorithm indicated that the patient had a colonic polyp, the colleague asked for an additional biopsy. It turned out that the result produced by the algorithm was correct (use the following scale: 1=have major doubts to 4=neutral to 7=fully believe).

Computer-aided characterization

- The second colleague reported that the machine learning algorithm is also capable of correctly classifying whether the colonic polyp was adenomatous or hyperplastic (use the following scale: 1=have major doubts to 4=neutral to 7=fully believe).

Computer-aided intervention

- Now suppose a third colleague told you that a machine learning algorithm can be applied to guide interventions. Endoscopists need a targeted biopsy from specific locations that harbor the lesion. The third colleague said that the algorithm can guide a biopsy needle more precisely than a human, using ultrasound imaging (use the following scale: 1=have major doubts to 4=neutral to 7=fully believe).

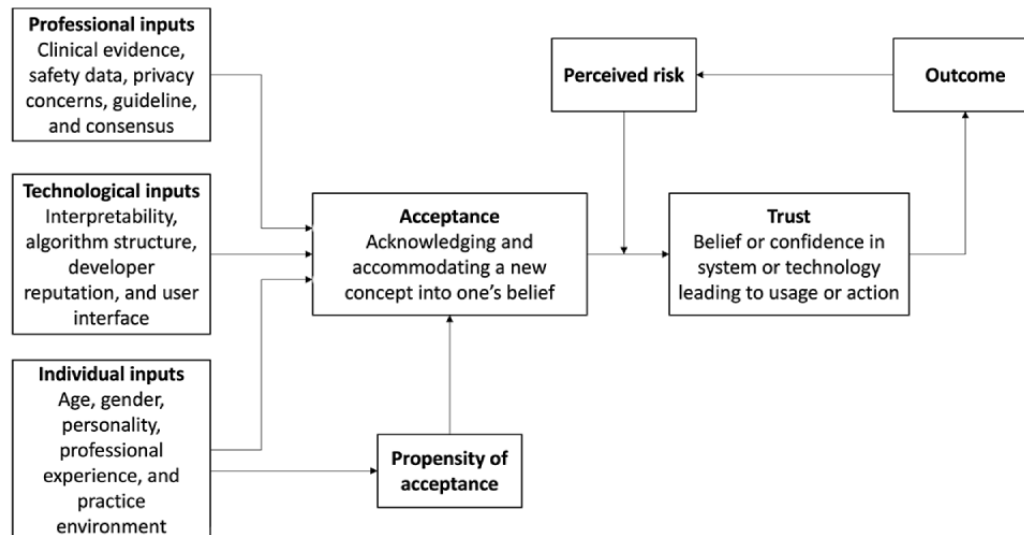
Table 1. Scenarios demonstrating AI^a use in gastroenterology practice from detection to characterization and intervention.

Scenario	Objective
Computer-aided detection: use of AI to assist in identifying the presence of colorectal polyps and improving adenoma detection rate.	To evaluate the acceptability of AI to assist in the interpretation of medical imagery in detecting colorectal lesions under different bowel preparations and colonic configurations
Computer-aided characterization: use of AI to classify whether a colonic polyp was adenomatous or hyperplastic.	To evaluate the acceptability of AI to differentiate (without histology) between adenoma (with variable degree of malignant potential) vs hyperplastic polyps (no malignant potent)
Computer-aided intervention: use of AI in an endoscopy to guide colonoscopic polypectomy.	To evaluate the acceptability of AI to decide which tool to use in assessing the completeness of polypectomy and risk of bleeding, perforation, or both.

^aAI: artificial intelligence.

Table 2. Survey items used to measure risk perception, acceptance, and trust.

Measure	Question text
Risk perception	I expect major risks involved with the artificial intelligence diagnosis.
Acceptance	Do you believe that machine learning algorithm can, in some cases (as in the one described above), better perform (the task, computer-aided detection, computer-aided characterization, Computer-aided intervention) than human beings?
Trust	I am ready to try the method myself

Figure 1. Conceptual model of perceived risk, acceptance, and trust on artificial intelligence decision aids.

Statistical Analysis

Statistically significant application pairs were identified by the Mann-Whitney U test (U test) or when there is dependence, the Wilcoxon signed-rank test (Wilcox test). Statistical significance is established at .05. Analyses were conducted in Python using the *scipy.stats* module (version 1.10.0; the SciPy community), *statsmodels* module (version 0.13.5), and the *Pingouin* statistical package (version 0.5.3) or SPSS (version 28; IBM Corp).

Correction for multiple testing was performed using Bonferroni correction, where the statistical threshold (α) was divided by the number of tests n , such that the adjusted P value threshold is given by α/n .

Power Analysis

Our hypothesis is that risk perception, acceptance, and trust will change according to the scenario (detection [CADE], characterization [CADx], or intervention [CADI]), with different levels of invasiveness. Based on an estimated effect size of 0.3 for trust, power, and risk perception with 0.95 power, we can calculate the minimum set of respondents needed to determine any significant differences of a given “size” in response to trust, risk perception, and acceptance measures across scenarios. Since every individual answers scenarios 1 to 3, the differences in the response of every individual can be estimated using a Wilcox test if we compare between pairs of scenarios. The required sample size to pick up a small-moderate effect size (based on Cohen d) of 0.3 with a power of 95% is 154. In this study, we have recruited 165 participants, and this should be enough to achieve sufficient statistical power.

Ethical Considerations

This study was approved by the Nanyang Technological University institutional review board (IRB-2022-756). Informed consent was obtained with ability to opt out. Data was anonymized, and no compensation was provided.

Results

Response and Nonresponse Bias

Tracking response rates can help determine the representativeness of a study, but due to the constraints of our institutional review board, we were not allowed to track individual respondents. During the initial phase of the study, we sent the survey to a distribution list of 151 participants with known dates. Applying an approximate 1-month window (October 21, 2022, to November 13, 2022), we obtained 128 responses. Thus, our estimated response rate is ~85% ($n=128$). While we were analyzing or cleaning up the data, we hoped to get more participants. In the subsequent weeks, we obtained 37 new responses. To compare early and late respondents, we aggregated the first 130 responses (collected between October 21, 2022, and December 29, 2022) as a single group to represent the early respondents and the remaining 35 (collected between January 10 to January 19, 2023) as the late responses. Comparing 130 early respondents against 35 late respondents using a Mann-Whitney U test with a Bonferroni-adjusted $\alpha=.0056$, we found no significant differences for risk, trust, and acceptance across each of the 3 scenarios. This suggests no significant difference between the early and late responses. The lowest obtained P value was .022 (trust in CADx), and the remaining P values were at least .30. Together, we take these

results as a proxy that nonresponder bias is not a strong concern. We also note the overall response rates are rather high; the survey was sent out to various gastroenterology associations as an open invitation, without individual follow-up. It is possible that AI is increasingly seen as transformative and important in the gastroenterology field, but there is not much work on understanding how perspectives on AI lead toward trust and adoption. Hence, invitees feel strongly about the matter and are more inclined to participate in this survey.

Unidimensionality and Reliability

Most items in our questionnaire were already used in other questionnaires and can be considered as validated. For the scenario-based questions used in this study, these are novel, as we needed to develop new instruments to explore new topics. Participants had to answer on three 7-point items (not at all to wholeheartedly) whether they accept, trust, and perceive risk on the method presented in each of the scenarios. Unidimensionality and reliability were verified and assured using confirmatory factor analysis and Omega Hierarchical, respectively (see [Multimedia Appendix 1](#) for details).

Cohort Characteristics

In total, 165 clinicians participated in the study. The survey completion rate was ~99.40% (n=165). Participants averaged 44.49 (SD 9.65) years, were mostly male (n=116, 70%), and predominantly specialized in gastroenterology (n=153, 92.72%; see [Table 3](#)).

The sample comprised gastroenterologists and GI surgeons with varied clinical experience: 93 (56.36%) participants have over 10 years' experience in practicing gastroenterology and 111 (66.81%) participants were consultants or senior consultants, mostly working in public hospitals (n=110, 66.67%). Most participants reported basic familiarity with AI (n=160, 96.97%; Q1: How familiar are you with AI?). Many were exposed at work, either directly (n=111, 67.27%; Q2: Have you ever used AI in your occupation?) or indirectly (n=112, 67.88%; Q6: Do you personally know other clinicians who use AI at work?).

Participants rated a mean score of 6.00 (SD 0.95) for intending to use AI when it becomes available in their workplace and a score of 5.50 (SD 1.24) for intending to use it to provide services to their patients. Participants rated a mean score of 5.83 (SD 1.37) for intention to use AI routinely in patient care. These figures suggest generally favorable attitudes toward adopting AI.

Table 3. Participant demographics and general characteristics.

Participant	Values (N=165), n (%)
Age (years), mean (SD) ^a	44.49 (9.65)
Gender^a	
Male	116 (75.32)
Female	38 (24.68)
Country or area^b	
Australia	3 (1.83)
Brunei Darussalam	7 (4.27)
Hong Kong	18 (10.98)
India	6 (3.66)
Indonesia	6 (3.66)
Japan	9 (5.49)
New Zealand	1 (0.61)
People's Republic of China	50 (30.49)
Philippines	1 (0.61)
Republic of Korea	2 (1.22)
Singapore	24 (14.63)
Taiwan	33 (20.12)
Main work setting^c	
Public hospital	110 (67.9)
Private hospital	28 (17.28)
Institute of higher learning	18 (11.11)
Community health center	1 (0.62)
Other	5 (3.09)
Current role at work^d	
Resident	19 (11.8)
Fellow	19 (11.8)
Consultant	57 (35.4)
Senior consultant	54 (33.54)
Other	12 (7.45)
Specialty^c	
Gastroenterology	153 (94.44)
Colorectal surgery	4 (2.47)
General surgery	2 (1.23)
Other	3 (1.85)
Practicing in specialty (years)^c	
Less than 5	39 (24.07)
5-10	30 (18.52)
11-20	48 (29.63)
Over 20	45 (27.78)

^a11 participants did not report their ages or gender.

^b1 participant did not report their country or area.

^c3 participants did not report their main work setting, specialty, and years practicing in a specialty.

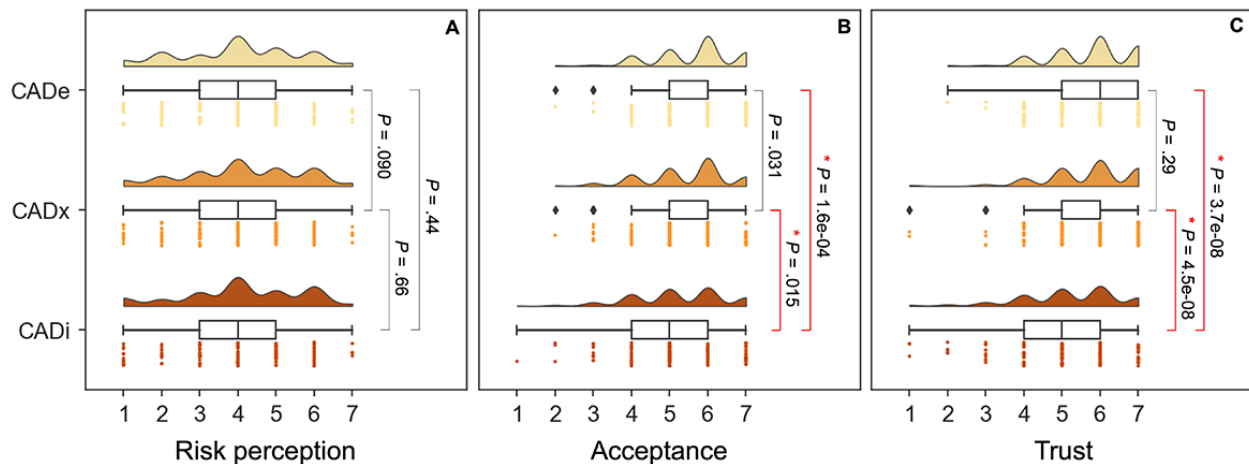
^d4 participants did not report their current role at work.

Scenario-Based Differentiation

When participants were exposed to three scenarios in medical practice that extend from (1) diagnosing and detecting colorectal polyps (CADE), (2) assessing the nature of pathology of polyps and predict risk of malignancy (CADx), and (3) adopting endoscopic or surgical intervention or removal of the polyps (CADi), clinicians expressed similar risk perceptions across all applications (Figure 2A: Median_{CADE}=Median_{CADx}=Median_{CADi}=4.0; Wilcoxon_{CADE-CADx}: $P=.09$; Wilcoxon_{CADE-CADi}: $P=.44$; Wilcoxon_{CADx-CADi}: $P=.66$).

However, there were clear application-specific differences in intention to accept AI in practice, with CADE and CADx rated higher than that of CADi (Figure 2B: Median_{CADE}=6.0, Median_{CADx}=6.0, Median_{CADi}=5.0; Wilcoxon_{CADE-CADx}: $P=.031$; Wilcoxon_{CADE-CADi}: $P=1.6 \times 10^{-4}$; Wilcoxon_{CADx-CADi}: $P=.02$). Similarly for trust, CADE and CADx were rated higher than CADi (Figure 2C: Median_{CADE}=6.0, Median_{CADx}=6.0, Median_{CADi}=5.0; Wilcoxon_{CADE-CADx}: $P=.29$; Wilcoxon_{CADE-CADi}: $P=3.7 \times 10^{-08}$; Wilcoxon_{CADx-CADi}: $P=4.5 \times 10^{-08}$).

Figure 2. Gastroenterologists' attitude toward using AI in the management of colorectal polyps: perceived risk, acceptance, and trust in 3 case scenarios of using AI-assisted colonoscopy in CADE, CADx, and adopting CADi with either surgery or endoscopy. Pairwise tests based on the Wilcoxon test were performed across scenarios. (A) Risk perception across CADE, CADx, and CADi applications. The raincloud plot comprises a 3-panel visualization with a density plot on top revealing density patterns, a box plot in the middle summarizing the median and IQR, and a univariate strip plot on the bottom showing the actual data distribution. No significant pairs were identified. (B) Acceptance across CADE, CADx, and CADi applications. Pairs with statistically significant differences are highlighted by a red connector and an asterisk. (C) Trust across CADE, CADx, and CADi applications. Pairs with statistically significant differences with a P value $\leq .02$ are highlighted by a red connector and an asterisk. AI: artificial intelligence; CADE: computer-aided detection; CADi: computer-aided intervention; CADx: computer-aided characterization.



Subgroup Analysis for Identification of Confounding Effects and Other Intrinsic Factors

We performed a subgroup analysis to investigate if factors such as gender, years of experience, and practice environment will affect risk perception, acceptance, and trust in AI for gastroenterology practice (Figure 3).

Male and female practitioners held similar risk perceptions. There was good concordance in their risk perception, acceptance, and trust toward using AI in gastroenterology practice (Figure 3A1, 3B1, and 3C1). Male participants tended to be less accepting and trusting, especially in CADi, although this difference is not statistically significant.

Next, we compared practitioners with 10 or less years of clinical experience ($n=69$) versus experienced practitioners with more than 10 years of clinical experience ($n=93$). While the overall trends of high acceptance and trust showed no difference between the 2 groups, experienced clinicians exhibited consistently lower risk perception than less experienced ones (Figure 3A2). This observation was statistically significant for all 3 scenarios (CADE: $P=9.7 \times 10^{-6}$; CADx: $P=1.7 \times 10^{-06}$; CADi:

$P=3.3 \times 10^{-04}$). We also compared practitioners of the rank senior consultant and consultant ($n=111$) against residents and fellows ($n=38$; Figure 3A3, 3B3, and 3C3). The acceptance and trust remained high, and the trend showed a good concordance between the 2 groups. A lower risk perception was found among senior consultants and consultants compared to residents and fellows (CADE: $P=.12$, CADx: $P=.10$, and CADi: $P=.27$). However, the difference is statistically insignificant. The years of experience in clinical practice appeared to have a stronger impact on risk perception than the rank held.

Finally, we compared practitioners from public hospitals with those from private hospitals (Figure 3A4, 3B4, and 3C4). There was no statistically significant difference between private hospital practitioners against their public counterparts, although there was a noticeable difference in CADx on acceptance (Figure 3B4). There was also a lower rate of acceptance and trust in using AI for intervention (CADi) compared to CADE and CADx. Despite not reaching statistical significance, we observed that the spread among private hospital respondents tended to exhibit greater variations. In some instances, the spread appeared to be

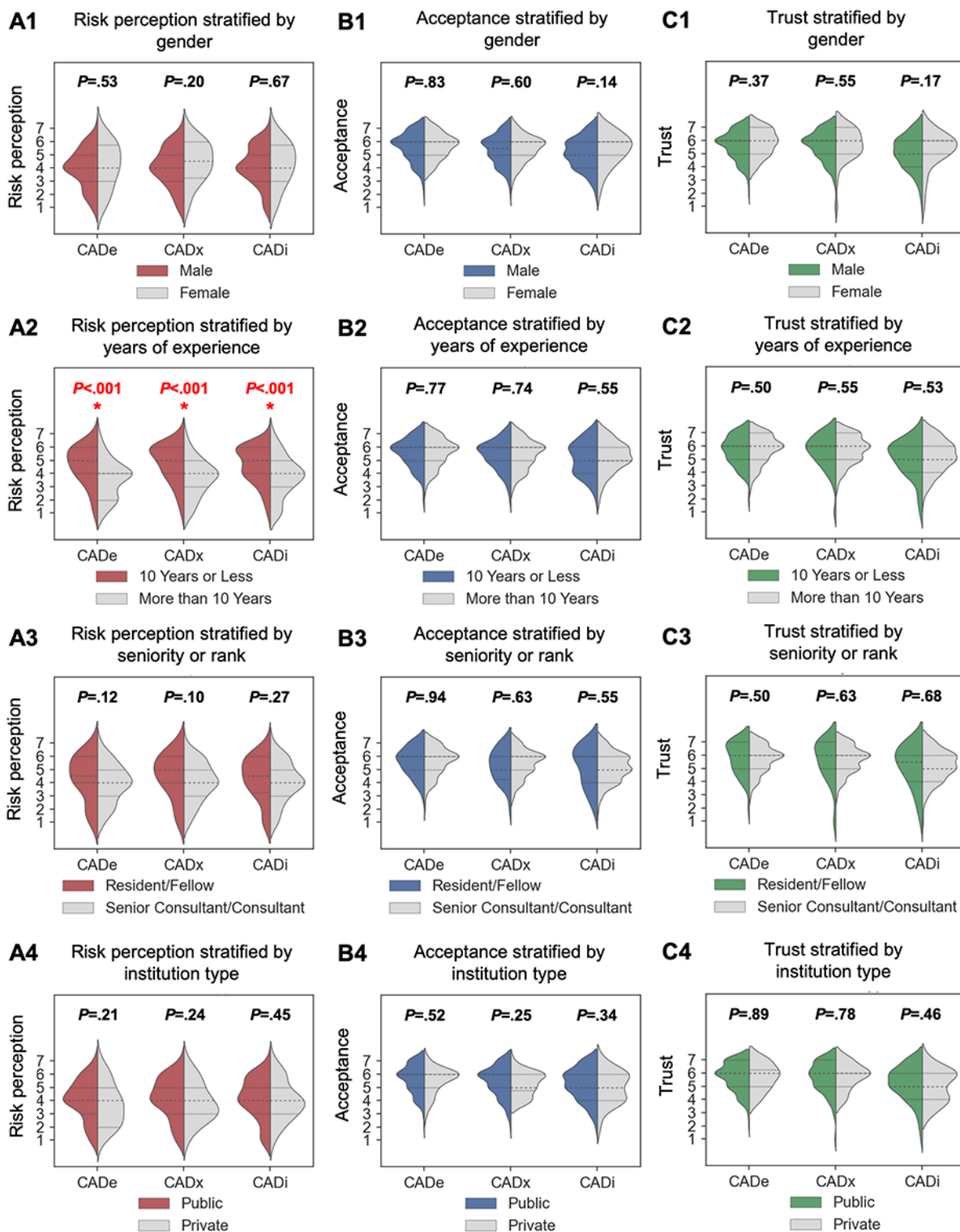
bimodal for CADi, suggesting that the private respondents could be a combination of 2 distinct subgroups.

The correlation among risk perception, acceptance, and trust was further analyzed by incorporating the years of experience of the participants by their years of practice in gastroenterology. In all 3 scenarios, there is a moderate correlation between acceptance and trust of AI in detecting polyps (CADe) and characterizing polyps (CADx). The influence of risk perception on acceptance and trust appears to be more diffused: noticeably, when trust and acceptance are both high, and it does not always coincide with low-risk perception.

We first used contingency tables combined with the Fisher exact test to evaluate the impact on the original relationships between trust and acceptance and after introducing risk perception (risk) as an interaction term. This was repeated for each scenario (CADx, CADi, and CADe; [Multimedia Appendix 1](#)). Using this approach, we find that after introducing risk perception, the distribution of values still largely follows that of the original data, suggesting that risk does not interact strongly with trust and acceptance. However, this does not mean that risk does not

influence these 2 factors. To further investigate, we performed a 2-way ANOVA to further study the influence of risk perception on acceptance and trust. The 2-way ANOVA revealed a statistically significant interaction in CADe ($F_{25}=3.37$; $P=1.6\times 10^{-05}$) but not in CADx ($F_{25}=1.40$; $P=.12$) and CADi ($F_{36}=1.35$; $P=.16$). Finally, we performed two sets of regression analyses with (1) acceptance and risk perception as independent variables and (2) acceptance, risk perception, and an interaction term that is the product of acceptance and risk perception ([Multimedia Appendix 1](#)). Acceptance had a statistically significant positive influence on trust for all 3 scenarios. Risk perception only has a statistically significant negative impact on trust for the first 2 scenarios (CADe and CADx). When we considered an interaction term, only CADe had a statistically significant impact on trust on all 3 terms. For CADx and CADi, this effect disappeared and only acceptance retained a statistically significant influence on trust. Thus, we believe risk perception has a weak association with trust and acceptance. Taken together, the relationship between trust, acceptance, and risk perception appears complex and is not straightforward.

Figure 3. Subgroup analysis of risk perception, acceptance, and trust stratified by year of experience, seniority (consultant+senior consultant vs fellow+resident), gender, and practicing environment (public vs private hospital). The visualization is a grouped violin plot with split violins. The left and right halves of the violin depict the distributions of 2 samples. If the 2 samples are similar, they will exhibit symmetry on both sides. The median lines for each sample have dashed lines, and these median lines are in turn, bordered by their respective 25th and 75th percentile lines depicted as dotted horizontal lines. Comparisons with statistically significant differences with P value $\leq .0014$ are flagged with a red asterisk. CADe: computer-aided detection; CADi: computer-aided intervention; CADx: computer-aided characterization.



Discussion

Principal Findings

The findings from our study demonstrate that gastroenterologists are generally familiar with AI and were frequently exposed to AI tools in medical settings. This may be because of the introduction of AI-assisted colonoscopy by various industries. In recent years, there are also numerous publications and seminars in the field of gastroenterology mentioning the success of using AI tools in diagnosis, risk prediction, and the treatment of GI conditions [23]. This suggests that they have a keen awareness of AI's future potential in clinical applications. However, our findings showed that acceptance is not an all-or-nothing choice, but the application or intention to use AI tools varied between different clinical scenarios as well as the nature and impact of AI participation.

When looking at scenario-specific acceptance and trust in AI, the responses vary. Our survey on AI use in detection (CADE), characterization (CADx), and intervention (CADi) of colonic polyps revealed wide acceptance disparity among practitioners (Figure 2). While CADE was more widely accepted, CADi was met with much greater resistance. The 3 AI scenarios that were presented to clinicians in this study varied in the degree of involvement a clinician has in certain procedures. Participants preferred CADi the least. These results agree with our hypothesis that trust, acceptance, and risk perception will change according to the scenario (detection [CADE], characterization [CADx], or intervention [CADi]), with different levels of invasiveness.

In this study, acceptance appeared to have little correlation with the perceived risk level of the procedures. Although certain case scenarios were considered by some as high risk, they do not necessarily warrant low acceptance or trust in using AI. Hence, the findings highlight the intricate relationship between the complexity of AI technologies and their acceptance. One intriguing finding is that participants with more (years of) experience appear to accept the risk and would trust the use of AI more than those who are less experienced. This probably indicates that they see the use of AI as an option or recommendation, instead as an obligation or necessity. Therefore, having more clinical experience may give clinicians greater confidence in their medical expertise and practice, thereby generating more confidence in risk mitigation when new technologies are introduced. Indeed, a study by Lawton et al [24] revealed more experienced doctors were much more at ease with uncertainty.

On the other hand, a general lack of AI familiarization and training in medical education may be one of the reasons that less experienced doctors perceive AI as more risky than regular or traditional practice. Chen et al [25] found that while most physicians and medical students were receptive to the use of AI, most also had concerns about the potential for unpredictable or incorrect results. The same study also stated that respondents were aware of AI's potential but lacked practical experience and related knowledge. Thus, introducing AI literacy and familiarization training early in medical careers may help mitigate risk aversion and promote responsible AI use in clinical

practice. Young doctors are also aware of their education gaps. In a study by Civaner et al [26], medical student respondents acknowledged a gap in “knowledge and skills related to AI applications” (96.2%), “applications for reducing medical errors” (95.8%), and “training to prevent and solve ethical problems that might arise as a result of using AI applications” (93.8%).

Our results suggest that although there is a moderate correlation between trust and acceptance, risk perception appeared invariant suggesting the relationship between trust and acceptance with risk perception is not straightforward and may implicate other factors and interactions than the relationships shown in Figure 1. Indeed, the invariance of risk perception across scenarios against acceptance suggests that there are other factors that influence the acceptance of AI (Figure 2). Among the tested factors, we find that risk acceptance is confounded with years of experience (Figure 3). Future studies should be conducted to better understand other drivers and barriers that influence acceptance, such as the perceived usefulness of using AI and whether AI tools may replace the jobs of clinicians in future practices. Qualitative studies, such as the use of focus group discussions, would also be useful to better understand clinicians' specific concerns in using AI and the impact of their concerns on the use of AI. Quantitatively, more complex data analysis methods may also be used in the future to understand the causal relationship between various factors and the acceptance of AI. As we proceed into deeper and larger cohort studies investigating trust and acceptance of AI, the development of powerful network methodologies can yield more insight. Indeed, simple statistical learning and even deep learning methods may soon become limited in their ability to explain complex and directed relationships among factors. We believe that causal analysis methods, such as Bayesian Belief Networks will soon become necessary and indispensable for explaining and modeling trust, acceptance, and risk perceptions on medical AI [27].

Limitations

There are limitations in this study. While this study provides invaluable insight into the Asia-Pacific region, we have only captured clinicians' perspectives despite there being other stakeholders whose voices and opinions matter. This includes nurses, endoscopy assistants, and patients. Future studies should aim to capture their perspectives and understand better how their opinions align or conflict with each other. This will help us navigate complex trust and acceptance issues more realistically and create valuable propositions and effective policies by adopting a multistakeholder perspective into consideration [28]. Participants in this study come from 5 countries with only 165 respondents. The generalizability of the findings can be strengthened by including more clinicians from different backgrounds and regions of practice. In future implementation studies, it may also be worthwhile to examine additional case scenarios such as the management of complicated inflammatory bowel diseases; choice of therapy for GI cancers and GI bleeding; and their corresponding trust, acceptance, and risk perceptions. This additional information will help us better contextualize how risk acceptance, acceptance, and trust change depending on practice.

Conclusions

This study is one of the first to examine risk perception, acceptance, and trust across different scenarios. It is one of the earliest reports of AI risk perception, acceptance, and trust among gastroenterologists, with a unique focus on the Asia-Pacific region. We found that gastroenterologists have, in general, a high acceptance and trust level of using AI-assisted

colonoscopy in the management of colorectal polyps. However, this level of trust depends on the application scenario. Moreover, the relationship among risk perception, acceptance, and trust in using AI in gastroenterology practice is not a straightforward correlation. Future studies are required to identify factors that influence the acceptance and trust of using AI in clinical practices.

Acknowledgments

This research or project is supported by the National Research Foundation, Singapore under its AI Singapore Programme (AISG3-GV-2021-009).

Conflicts of Interest

None declared.

Multimedia Appendix 1

Survey questions and supplementary results.

[[DOCX File, 145 KB - ai_v3i1e50525_appl.docx](#)]

References

1. Elmore JG, Lee CI. Artificial intelligence in medical imaging—learning from past mistakes in mammography. *JAMA Health Forum* 2022;3(2):e215207 [FREE Full text] [doi: [10.1001/jamahealthforum.2021.5207](https://doi.org/10.1001/jamahealthforum.2021.5207)] [Medline: [36218833](https://pubmed.ncbi.nlm.nih.gov/36218833/)]
2. Ren Y, Loftus TJ, Datta S, Ruppert MM, Guan Z, Miao S, et al. Performance of a machine learning algorithm using electronic health record data to predict postoperative complications and report on a mobile platform. *JAMA Netw Open* 2022;5(5):e2211973 [FREE Full text] [doi: [10.1001/jamanetworkopen.2022.11973](https://doi.org/10.1001/jamanetworkopen.2022.11973)] [Medline: [35576007](https://pubmed.ncbi.nlm.nih.gov/35576007/)]
3. Hoiland RL, Rikhray KJK, Thiara S, Fordyce C, Kramer AH, Skrifvars MB, et al. Neurologic prognostication after cardiac arrest using brain biomarkers: a systematic review and meta-analysis. *JAMA Neurol* 2022;79(4):390-398 [FREE Full text] [doi: [10.1001/jamaneurol.2021.5598](https://doi.org/10.1001/jamaneurol.2021.5598)] [Medline: [35226054](https://pubmed.ncbi.nlm.nih.gov/35226054/)]
4. Piette JD, Newman S, Krein SL, Marinec N, Chen J, Williams DA, et al. Patient-centered pain care using artificial intelligence and mobile health tools: a randomized comparative effectiveness trial. *JAMA Intern Med* 2022;182(9):975-983 [FREE Full text] [doi: [10.1001/jamainternmed.2022.3178](https://doi.org/10.1001/jamainternmed.2022.3178)] [Medline: [35939288](https://pubmed.ncbi.nlm.nih.gov/35939288/)]
5. Loftus TJ, Tighe PJ, Filiberto AC, Efron PA, Brakenridge SC, Mohr AM, et al. Artificial intelligence and surgical decision-making. *JAMA Surg* 2020;155(2):148-158 [FREE Full text] [doi: [10.1001/jamasurg.2019.4917](https://doi.org/10.1001/jamasurg.2019.4917)] [Medline: [31825465](https://pubmed.ncbi.nlm.nih.gov/31825465/)]
6. Asan O, Bayrak AE, Choudhury A. Artificial intelligence and human trust in healthcare: focus on clinicians. *J Med Internet Res* 2020;22(6):e15154 [FREE Full text] [doi: [10.2196/15154](https://doi.org/10.2196/15154)] [Medline: [32558657](https://pubmed.ncbi.nlm.nih.gov/32558657/)]
7. Schepart A, Burton A, Durkin L, Fuller A, Charap E, Bhambri R, et al. Artificial intelligence-enabled tools in cardiovascular medicine: a survey of current use, perceptions, and challenges. *Cardiovasc Digit Health J* 2023;4(3):101-110 [FREE Full text] [doi: [10.1016/j.cvdhj.2023.04.003](https://doi.org/10.1016/j.cvdhj.2023.04.003)] [Medline: [37351333](https://pubmed.ncbi.nlm.nih.gov/37351333/)]
8. Frank DA, Elbæk CT, Børsting CK, Mitkidis P, Otterbring T, Borau S. Drivers and social implications of artificial intelligence adoption in healthcare during the COVID-19 pandemic. *PLoS One* 2021;16(11):e0259928 [FREE Full text] [doi: [10.1371/journal.pone.0259928](https://doi.org/10.1371/journal.pone.0259928)] [Medline: [34807907](https://pubmed.ncbi.nlm.nih.gov/34807907/)]
9. Alraja MN, Farooque MMJ, Khashab B. The effect of security, privacy, familiarity, and trust on users' attitudes toward the use of the IoT-based healthcare: the mediation role of risk perception. *IEEE Access* 2019;7:111341-111354 [FREE Full text] [doi: [10.1109/access.2019.2904006](https://doi.org/10.1109/access.2019.2904006)]
10. Choudhury A, Asan O, Medow JE. Effect of risk, expectancy, and trust on clinicians' intent to use an artificial intelligence system—blood utilization calculator. *Appl Ergon* 2022;101:103708. [doi: [10.1016/j.apergo.2022.103708](https://doi.org/10.1016/j.apergo.2022.103708)] [Medline: [35149301](https://pubmed.ncbi.nlm.nih.gov/35149301/)]
11. Mori Y, Neumann H, Misawa M, Kudo SE, Bretthauer M. Artificial intelligence in colonoscopy—now on the market. What's next? *J Gastroenterol Hepatol* 2021;36(1):7-11. [doi: [10.1111/jgh.15339](https://doi.org/10.1111/jgh.15339)] [Medline: [33179322](https://pubmed.ncbi.nlm.nih.gov/33179322/)]
12. Walradt T, Berzin TM. Artificial intelligence in gastroenterology. In: Greenhill AT, Chahal D, Byrne MF, Parsa N, Ahmad O, Bagci U, editors. *AI in Clinical Medicine*. Hoboken, NJ: Wiley; 2023:176-183.
13. van der Zander QEW, van der Ende-van Loon MCM, Janssen JMM, Winkens B, van der Sommen F, Masclee AAM, et al. Artificial intelligence in (gastrointestinal) healthcare: patients' and physicians' perspectives. *Sci Rep* 2022;12(1):16779 [FREE Full text] [doi: [10.1038/s41598-022-20958-2](https://doi.org/10.1038/s41598-022-20958-2)] [Medline: [36202957](https://pubmed.ncbi.nlm.nih.gov/36202957/)]

14. Ye T, Xue J, He M, Gu J, Lin H, Xu B, et al. Psychosocial factors affecting artificial intelligence adoption in health care in China: cross-sectional study. *J Med Internet Res* 2019;21(10):e14316 [FREE Full text] [doi: [10.2196/14316](https://doi.org/10.2196/14316)] [Medline: [31625950](https://pubmed.ncbi.nlm.nih.gov/31625950/)]
15. Gupta S, Kamboj S, Bag S. Role of risks in the development of responsible artificial intelligence in the digital healthcare domain. *Inf Syst Front* 2021;25(6):2257-2274. [doi: [10.1007/s10796-021-10174-0](https://doi.org/10.1007/s10796-021-10174-0)]
16. Ajzen I. The theory of planned behavior. *Organ Behav Hum Decis Process* 1991;50(2):179-211. [doi: [10.1016/0749-5978\(91\)90020-t](https://doi.org/10.1016/0749-5978(91)90020-t)]
17. Davis FD. A technology acceptance model for empirically testing new end-user information systems: theory and results. Massachusetts Institute of Technology. 1985. URL: <https://dspace.mit.edu/handle/1721.1/15192> [accessed 2024-01-09]
18. Kader R, Baggaley RF, Hussein M, Ahmad OF, Patel N, Corbett G, et al. Survey on the perceptions of UK gastroenterologists and endoscopists to artificial intelligence. *Frontline Gastroenterol* 2022;13(5):423-429 [FREE Full text] [doi: [10.1136/flgastro-2021-101994](https://doi.org/10.1136/flgastro-2021-101994)] [Medline: [36046492](https://pubmed.ncbi.nlm.nih.gov/36046492/)]
19. Hah H, Goldin DS. How clinicians perceive artificial intelligence-assisted technologies in diagnostic decision making: mixed methods approach. *J Med Internet Res* 2021;23(12):e33540 [FREE Full text] [doi: [10.2196/33540](https://doi.org/10.2196/33540)] [Medline: [34924356](https://pubmed.ncbi.nlm.nih.gov/34924356/)]
20. Mayer RC, Davis JH, Schoorman FD. An integrative model of organizational trust. *Acad Manage Rev* 1995;20(3):709-734 [FREE Full text] [doi: [10.5465/amr.1995.9508080335](https://doi.org/10.5465/amr.1995.9508080335)]
21. Schoorman FD, Mayer RC, Davis JH. An integrative model of organizational trust: past, present, and future. *Acad Manage Rev* 2007;32(2):344-354. [doi: [10.5465/amr.2007.24348410](https://doi.org/10.5465/amr.2007.24348410)]
22. Solberg E, Kaarstad M, Eitrheim MHR, Bisio R, Reegård K, Bloch M. A conceptual model of trust, perceived risk, and reliance on AI decision aids. *Group Organ Manage* 2022;47(2):187-222. [doi: [10.1177/10596011221081238](https://doi.org/10.1177/10596011221081238)]
23. Kröner PT, Engels MM, Glicksberg BS, Johnson KW, Mzaik O, van Hooft JE, et al. Artificial intelligence in gastroenterology: a state-of-the-art review. *World J Gastroenterol* 2021;27(40):6794-6824 [FREE Full text] [doi: [10.3748/wjg.v27.i40.6794](https://doi.org/10.3748/wjg.v27.i40.6794)] [Medline: [34790008](https://pubmed.ncbi.nlm.nih.gov/34790008/)]
24. Lawton R, Robinson O, Harrison R, Mason S, Conner M, Wilson B. Are more experienced clinicians better able to tolerate uncertainty and manage risks? a vignette study of doctors in three NHS emergency departments in England. *BMJ Qual Saf* 2019;28(5):382-388 [FREE Full text] [doi: [10.1136/bmjqs-2018-008390](https://doi.org/10.1136/bmjqs-2018-008390)] [Medline: [30728187](https://pubmed.ncbi.nlm.nih.gov/30728187/)]
25. Chen M, Zhang B, Cai Z, Seery S, Gonzalez MJ, Ali NM, et al. Acceptance of clinical artificial intelligence among physicians and medical students: a systematic review with cross-sectional survey. *Front Med (Lausanne)* 2022;9:990604 [FREE Full text] [doi: [10.3389/fmed.2022.990604](https://doi.org/10.3389/fmed.2022.990604)] [Medline: [36117979](https://pubmed.ncbi.nlm.nih.gov/36117979/)]
26. Civaner MM, Uncu Y, Bulut F, Chalil EG, Tatli A. Artificial intelligence in medical education: a cross-sectional needs assessment. *BMC Med Educ* 2022;22(1):772 [FREE Full text] [doi: [10.1186/s12909-022-03852-3](https://doi.org/10.1186/s12909-022-03852-3)] [Medline: [36352431](https://pubmed.ncbi.nlm.nih.gov/36352431/)]
27. Vasudevan RK, Ziatdinov M, Vlcek L, Kalinin SV. Off-the-shelf deep learning is not enough, and requires parsimony, Bayesianity, and causality. *npj Comput Mater* 2021;7(1):16 [FREE Full text] [doi: [10.1038/s41524-020-00487-0](https://doi.org/10.1038/s41524-020-00487-0)]
28. Güngör H. Creating value with artificial intelligence: a multi-stakeholder perspective. *J Creat Value* 2020;6(1):72-85 [FREE Full text] [doi: [10.1177/2394964320921071](https://doi.org/10.1177/2394964320921071)]

Abbreviations

- AI:** artificial intelligence
CADe: computer-aided detection
CADi: computer-aided intervention
CADx: computer-aided characterization
GI: gastrointestinal

Edited by K El Emam, B Malin; submitted 05.07.23; peer-reviewed by D Lungu, Z Li; comments to author 01.08.23; revised version received 28.08.23; accepted 23.11.23; published 07.03.24.

Please cite as:

Goh WWB, Chia KYA, Cheung MFK, Kee KM, Lwin MO, Schulz PJ, Chen M, Wu K, Ng SSM, Lui R, Ang TL, Yeoh KG, Chiu HM, Wu DC, Sung JJY

Risk Perception, Acceptance, and Trust of Using AI in Gastroenterology Practice in the Asia-Pacific Region: Web-Based Survey Study
JMIR AI 2024;3:e50525

URL: <https://ai.jmir.org/2024/1/e50525>

doi: [10.2196/50525](https://doi.org/10.2196/50525)

PMID: [38875591](https://pubmed.ncbi.nlm.nih.gov/38875591/)

©Wilson WB Goh, Kendrick YA Chia, Max FK Cheung, Kalya M Kee, May O Lwin, Peter J Schulz, Minhu Chen, Kaichun Wu, Simon SM Ng, Rashid Lui, Tiing Leong Ang, Khay Guan Yeoh, Han-mo Chiu, Deng-chyang Wu, Joseph JY Sung. Originally published in JMIR AI (<https://ai.jmir.org>), 07.03.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR AI, is properly cited. The complete bibliographic information, a link to the original publication on <https://www.ai.jmir.org/>, as well as this copyright and license information must be included.

Original Paper

Assessment of ChatGPT-3.5's Knowledge in Oncology: Comparative Study with ASCO-SEP Benchmarks

Roupen Odabashian^{1*}, MD; Donald Bastin^{2*}, MD; Georden Jones^{3*}, PhD; Maria Manzoor, MD; Sina Tangestaniapour, MD; Malke Assad⁴, MD; Sunita Lakhani⁵, MD; Maritsa Odabashian^{3,6}, Bsc (c); Sharon McGee^{7,8*}, MD, PhD

¹Department of Oncology, Barbara Ann Karmanos Cancer Institute, Wayne State University, Detroit, MI, United States

²Department of Medicine, Division of Internal Medicine, The Ottawa Hospital and the University of Ottawa, Ottawa, ON, Canada

³Mary A Rackham Institute, University of Michigan, Ann Arbor, MI, United States

⁴Department of Plastic Surgery, University of Pittsburgh Medical Center, Pittsburgh, PA, United States

⁵Department of Medicine, Division of Internal Medicine, Jefferson Abington Hospital, Philadelphia, PA, United States

⁶The Ottawa Hospital Research Institute, Ottawa, ON, Canada

⁷Department of Medicine, Division of Medical Oncology, The Ottawa Hospital and the University of Ottawa, Ottawa, ON, Canada

⁸Cancer Therapeutics Program, Ottawa Hospital Research Institute, Ottawa, ON, Canada

*these authors contributed equally

Corresponding Author:

Roupen Odabashian, MD

Department of Oncology, Barbara Ann Karmanos Cancer Institute, Wayne State University

4100 John R St

Detroit, MI, 48201

United States

Phone: 1 (313) 745 3000 ext 7731

Fax: 1 313 576 8120

Email: roupen.odabashian@mclaren.org

Abstract

Background: ChatGPT (Open AI) is a state-of-the-art large language model that uses artificial intelligence (AI) to address questions across diverse topics. The American Society of Clinical Oncology Self-Evaluation Program (ASCO-SEP) created a comprehensive educational program to help physicians keep up to date with the many rapid advances in the field. The question bank consists of multiple choice questions addressing the many facets of cancer care, including diagnosis, treatment, and supportive care. As ChatGPT applications rapidly expand, it becomes vital to ascertain if the knowledge of ChatGPT-3.5 matches the established standards that oncologists are recommended to follow.

Objective: This study aims to evaluate whether ChatGPT-3.5's knowledge aligns with the established benchmarks that oncologists are expected to adhere to. This will furnish us with a deeper understanding of the potential applications of this tool as a support for clinical decision-making.

Methods: We conducted a systematic assessment of the performance of ChatGPT-3.5 on the ASCO-SEP, the leading educational and assessment tool for medical oncologists in training and practice. Over 1000 multiple choice questions covering the spectrum of cancer care were extracted. Questions were categorized by cancer type or discipline, with subcategorization as treatment, diagnosis, or other. Answers were scored as correct if ChatGPT-3.5 selected the answer as defined by ASCO-SEP.

Results: Overall, ChatGPT-3.5 achieved a score of 56.1% (583/1040) for the correct answers provided. The program demonstrated varying levels of accuracy across cancer types or disciplines. The highest accuracy was observed in questions related to developmental therapeutics (8/10; 80% correct), while the lowest accuracy was observed in questions related to gastrointestinal cancer (102/209; 48.8% correct). There was no significant difference in the program's performance across the predefined subcategories of diagnosis, treatment, and other ($P=.16$, which is greater than .05).

Conclusions: This study evaluated ChatGPT-3.5's oncology knowledge using the ASCO-SEP, aiming to address uncertainties regarding AI tools like ChatGPT in clinical decision-making. Our findings suggest that while ChatGPT-3.5 offers a hopeful outlook for AI in oncology, its present performance in ASCO-SEP tests necessitates further refinement to reach the requisite competency levels. Future assessments could explore ChatGPT's clinical decision support capabilities with real-world clinical

scenarios, its ease of integration into medical workflows, and its potential to foster interdisciplinary collaboration and patient engagement in health care settings.

(JMIR AI 2024;3:e50442) doi:[10.2196/50442](https://doi.org/10.2196/50442)

KEYWORDS

artificial intelligence; ChatGPT-3.5; language model; medical oncology

Introduction

OpenAI released ChatGPT, a pioneering artificial intelligence (AI) language model, in late 2022. ChatGPT-3 is an AI chatbot that can comprehend user input and react to it in a manner that is natural and human-like [1]. The program was trained on a large body of data sourced from the internet, including textbooks, articles, social media posts, and web-based forums, up to the last quarter of 2021 [2]. It works by analyzing user input text to generate a response using a probabilistic distribution of words and phrases derived from its training data. To date, it has significantly impacted numerous disciplines, including law, health care, and medical education [3-6]. Large language models like ChatGPT-3.5 represent a significant advancement in the preceding class of deep learning-based models, by facilitating the interpretation, processing, and production of natural language [7].

The use of AI has rapidly emerged as a promising approach in the health care industry, where it has been applied to medical imaging analysis, drug discovery, and patient monitoring [8]. Recent research has evaluated ChatGPT-3.5's abilities to respond to standardized questions from professional examinations for law and the United States Medical Licensing Examination (USMLE) [3,4]. ChatGPT-3.5 was able to achieve passing grades on these examinations while providing logical and informative explanations. Additionally, studies have been conducted to assess ChatGPT's capabilities in responding to international medical licensing examinations from countries such as Italy, France, Spain, the United Kingdom, and India. The success rates observed ranged between 22% and 73% [9].

AI and Chat GPT showcase substantial promise in augmenting medical consultations, offering preliminary diagnostic suggestions, and providing a vast knowledge base for medical practitioners and patients alike [10]. However, while it embarks on a path toward a more integrated health care AI system, several limitations hinder its full potential. The model's reliance on historical data without the ability to access real-time patient data can lead to outdated or inaccurate information dissemination. Additionally, its inability to comprehend nuanced human emotions and the ethical implications surrounding patient data privacy remain significant hurdles [11].

AI has displayed a notable deficiency in grasping context and nuance, elements that are fundamental for delivering safe and effective patient care [12]. Furthermore, analyzing the prospects of job automation in health care, Frey and Osborne [13] have projected that while administrative roles within the sector, such as health information technicians, exhibit a high likelihood of automation at 91%, the odds plummet to a mere 0.42% for the

automation of roles held by physicians and surgeons. This stark contrast underscores the intricate nature of medical practice, which extends beyond the mere application of codified knowledge. Additionally, there is a burgeoning discussion around the ethical dimensions of using conversational AI in medical practice. The crux of the issue revolves around the substantial volume of high-quality data required to train these models. Present-day algorithms are often honed on biased data sets, inheriting not just the availability, selection, and confirmation biases inherent in the data but also displaying a propensity to exacerbate these biases [14]. Looking ahead, the evolving capabilities of AI hint at the potential for tackling more sophisticated tasks, such as orchestrating experiments or future clinical trials [15] or engaging in peer review processes [16].

The American Society of Clinical Oncology Self-Evaluation (ASCO-SEP) program created a comprehensive educational program to help physicians keep up to date with the many rapid advances in the field. The question bank consists of multiple choice questions (MCQs) addressing the many facets of cancer care, including diagnosis, treatment, and supportive care. It is intended to evaluate participants' knowledge and give them feedback to direct future learning. The program is largely regarded as the leading resource for cancer specialists seeking to gain and maintain professional licensure in the field of medical oncology [17].

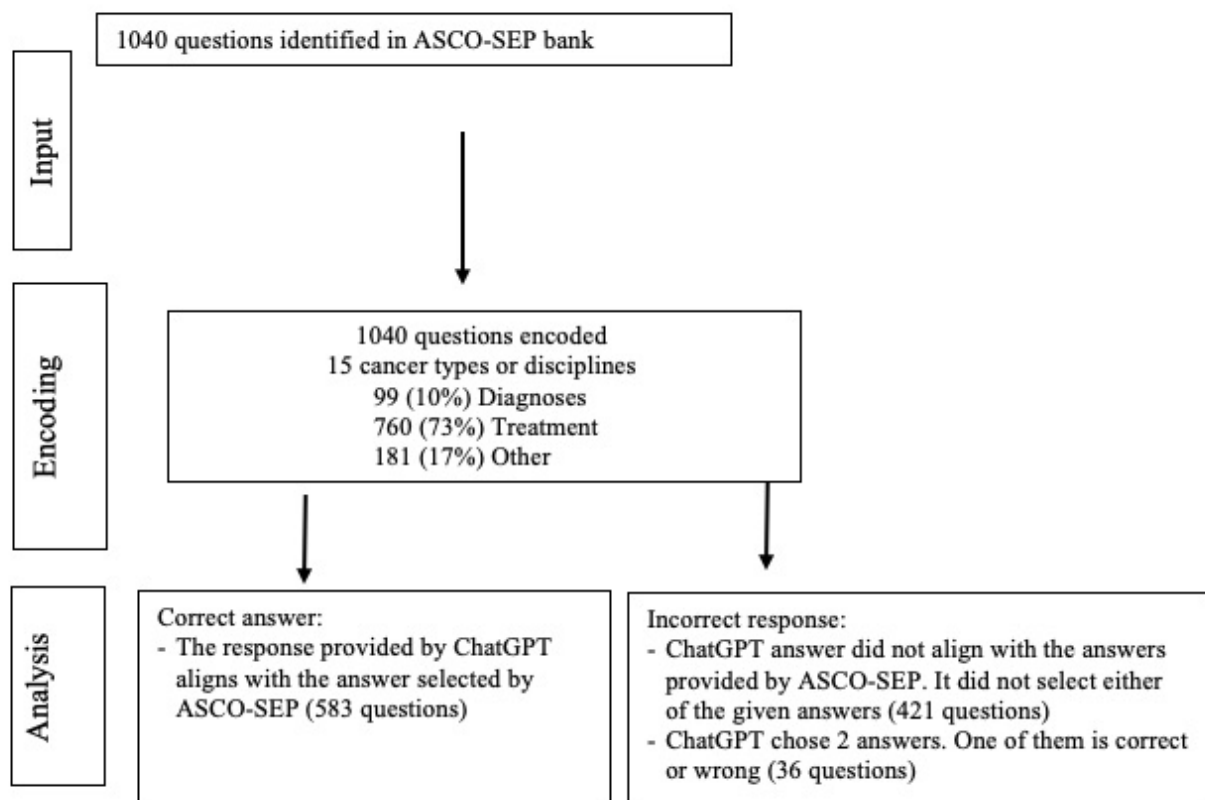
However, the evolving complexities of oncological care necessitate additional tools that can aid oncologists in clinical decision-making. By assessing ChatGPT-3's ability to answer ASCO-SEP questions, this study's objective is to understand ChatGPT's potential to serve as a supportive instrument in clinical decisions, offering instantaneous insights for health care providers, and to identify novel and efficient educational aids in oncology, with a specific emphasis on their role in clinical decisions.

Methods

Input Data

Questions were sourced from ASCO-SEP, which consists of approximately 1000 MCQs covering the spectrum of cancer care. The question bank was accessed from February 2023 to March 2023. As ChatGPT-3.5 can only generate responses to textual data, the study excluded questions with images, tables, or other non-textual content. Questions consisted of an information stem followed by a specific question with 3-5 possible answers (A-E), along with their corresponding letter choices, only 1 of which was correct. Figure 1 illustrates the workflow for data sourcing, input, encoding, and analysis.

Figure 1. Schematic of sourcing, encoding, and scoring procedures. ASCO-SEP: American Society of Clinical Oncology Self-Evaluation Program.



Before proceeding with the analysis, a random spot check was performed. For this, a random subset of the ASCO-SEP questions was selected, and their answers, explanations, or related content were manually cross-referenced with Google's index to ensure that they were not present before January 1, 2022, the last date accessible to the ChatGPT training data.

During this study, we used the free version of ChatGPT-3.5. At that time, ChatGPT-4 and its associated plugins were not yet available.

Encoding

We imported individual ASCO-SEP questions, including the information stem and multiple-choice response options, into the ChatGPT-3.5 interface. The questions were formatted to include the question stem, followed by each potential response on a separate line. We did not change the structure of the questions given to ChatGPT-3.5 and entered them in the original format provided by ASCO-SEP without altering the phrasing or the wording. A new conversation session was started in ChatGPT for each question. We did not provide ChatGPT-3.5 with any prompts and offered only one opportunity to answer each question.

Data Analysis

Selected questions were grouped by cancer type or discipline (eg, breast, lung, and colon cancer) with further

subcategorization based on content such as treatment, diagnosis, or other. ChatGPT was deemed to have responded correctly if it chose the correct answer as defined and provided by ASCO-SEP. The study team did not define or determine the correct answer. The program was not asked to provide justifications or references for answers. No point was assigned if ChatGPT-3.5 provided an answer that was not from the options given. Questions where ChatGPT-3.5 chose 2 possible answers or chose multiple answers and did not commit to a single best answer were also considered wrong, even if 1 of the responses was correct.

For statistical analysis, data were logged, scored, and analyzed in Excel (Microsoft Corp). Specifically, a chi-square test was performed to determine if there was a significant difference in the distribution of correct answers across different categories or groups.

Results

A total of 1040 questions were extracted from the ASCO-SEP question bank.

The questions covered 15 cancer types or disciplines. The largest portion focused on breast (223/1040; 21.4%) and gastrointestinal (209/1040; 20%) cancers, with $\leq 1\%$ (13/1040) covering central nervous system malignancies, developmental therapeutics, and prevention/epidemiology (Table 1).

Table 1. Question distribution and proportions by cancer type or specialty area.

Cancer type or discipline	Number of questions (N=1040), n (%)
Breast cancer	223 (21.4)
Gastrointestinal cancer	209 (20)
Thoracic oncology	137 (13.1)
Hematological malignancies	121 (11.6)
Genitourinary cancer	97 (9)
Melanoma and skin cancer	43 (4)
Sarcoma	36 (3)
Head and neck	36 (3)
Gynecologic cancers	36 (3)
General oncology	29 (3)
Supportive and palliative care	28 (3)
Genetics and genomics	17 (2)
Central nervous system	13 (1)
Developmental therapeutics	10 (1)
Prevention and epidemiology	5 (0.5)

Varying levels of accuracy were observed in ChatGPT-3.5's performance in answering questions based on different cancer types or disciplines (Table 2). The highest accuracy was achieved in questions related to developmental therapeutics (8/10; 80% correct), while the lowest accuracy was observed for questions related to gastrointestinal cancer (102/209; 48.8% correct).

Table 2. Accuracy rates by cancer type or specialty area.

Cancer type or discipline	Discipline-specific accuracy rates, n/N (%)
Developmental therapeutics	8/10 (80)
Central nervous system	10/13 (77)
Melanoma and skin cancer	28/43 (65)
Genetics and genomics	11/17 (65)
General oncology	18/29 (62)
Gynecologic cancers	22/36 (61)
Supportive and palliative care	17/28 (61)
Prevention and epidemiology	3/5 (60)
Head and neck	21/36 (58)
Breast cancer	130/223 (58.3)
Sarcoma	20/36 (57)
Thoracic oncology	77/137 (56)
Hematological malignancies	66/121 (55)
Genitourinary cancer	49/97 (51)
Gastrointestinal cancer	102/209 (48.8)
Total	583/1040 (56.1)

Questions were further subcategorized as "diagnosis," "treatment," and "other," with the latter covering topics such as biostatistics, cancer staging, and treatment complications. Out of the total questions, 73.1% (760/1040) were related to cancer treatment, 10% (99/1040) focused on diagnosis, and the remaining 17.4% (181/1040) were categorized as "other" (Table

3). Accuracy based on subcategory also varied, with 55% (418/760) of treatment questions, 63% (62/99) of diagnosis questions, and 56.9% (103/181) of "other" questions answered correctly (Table 2). There was no significant difference in the program's performance across the predefined subcategories of

diagnosis, treatment, and other ($P=.16$, which is greater than $.05$).

Table 3. ChatGPT-3.5 performance on questions per subcategory.

Category	Number of questions, n (%)	Overall accuracy, n/N (%)	<i>P</i> value ^a
Treatment	760 (73.1)	418/760 (55)	.16
Diagnosis	99 (10)	62/99 (63)	.16
Other	181 (17.4)	103/181 (56.9)	.16
Overall	1040 (100)	583/1040 (56.3)	.16

^aChi-square test.

Overall, ChatGPT-3.5 achieved a score of 56.3% (583/1040) for correct answers provided across all categories. Of note, responses were marked as incorrect if ChatGPT-3.5 provided 2 or more answers, even if 1 of those answers was correct (37/1040, 3%; [Figure 1](#)).

Discussion

Overview

In this study, we evaluated the performance of ChatGPT-3.5 in answering ASCO-SEP questions designed for medical oncologists in training and practice to support licensure and ongoing medical education. To facilitate a fair and rigorous assessment, spot checks were performed to ensure answers were not present in the program training data, and questions were entered in separate sessions to avoid grounding bias. Furthermore, questions were presented in their original format, as seen by physicians, with no changes made to prompt the program.

Over 1000 questions were posed to the program, spanning the spectrum of cancer care, with an overall score of 56.3% (583/1040) achieved. While promising, this is, however, below the accepted threshold of 70% that is required by ASCO-SEP to claim CME credits using their question bank [18].

Since the launch of ChatGPT-3.5, several studies have evaluated the program's performance on medical examinations. A notable study conducted by Kung et al [3] assessed ChatGPT-3.5's performance on the USMLE taken by US medical students. The results showed that ChatGPT-3.5 performed at, or near, the passing threshold for all 3 examinations. Specifically, the accuracy rates for USMLE Steps 1, 2 CK, and 3 were 68.0%, 58.3%, and 62.4%, respectively, which are acceptable passing scores. Gilson et al [19] reported similar results, where ChatGPT-3.5 scored 60% on USMLE test questions. It is worth noting that although the authors used questions published on the USMLE website after the training date cutoff for ChatGPT, which is late 2021, many of these questions were similar to those published in previous years. Moreover, these questions were discussed on web-based forums, which may explain the higher scores achieved [20]. Additionally, previous studies have evaluated ChatGPT-3.5's performance in microbiology [21] and pathology [22] and have shown promising outcomes in these fields with an accuracy rate of 80%.

Several factors might explain why ChatGPT's performs differently on USMLE compared to ASCO-SEP questions. First,

the ASCO-SEP is tailored for medical oncologists, delving deep into cancer care, while USMLE caters to a broader set of medical students, covering general medical knowledge. Given that ChatGPT-3.5's training data spans a wide range of topics, it's plausible that the content aligns more with the generalized medical queries of USMLE than the specialized focus of ASCO-SEP. Additionally, the structure and phrasing of questions play a critical role, potentially influencing AI's response accuracy. The questions within the USMLE typically features keywords that assist students in selecting an answer from the provided options. Conversely, the ASCO-SEP presents more specialized questions, challenging physicians' ability to discern first- and second-line treatments for a specified condition [23]. For instance, in 1 of the numerous subreddits [24] available web-based that was likely included in ChatGPT's training data set [25] students discuss how certain keywords aid them in answering examination questions. These data might have assisted ChatGPT in responding to USMLE questions in a previous paper that tested ChatGPT's performance on the USMLE [3,19]. However, such keywords are not used or discussed among physicians engaging with ASCO-SEP questions.

There are additional possible explanations for the observed performance of ChatGPT-3.5 in this study. One key factor is the comprehensive data set of over 1000 questions used, which allowed for a more thorough and holistic evaluation of the program's performance compared to previous studies [3,19,26,27]. Another contributing factor may be the dynamic and rapid scientific and clinical advances that occur in the field of oncology, which ChatGPT-3.5 could not fully tackle given that its training data is limited to pre-2022 internet data, with restricted access to key databases in the field like PubMed [28].

ChatGPT-3.5 demonstrated varying levels of accuracy in answering questions across the different cancer types and disciplines. Questions related to developmental therapeutics had the highest accuracy rate (80%, 8/10); however, the limited question sample size may not have allowed a complete assessment. Indeed, ChatGPT-3.5's lowest score was achieved in gastrointestinal cancer, which contained one of the largest numbers of questions in the bank (102/209, 48.8%), suggesting that broader assessments may identify more knowledge gaps. This study did not, however, find any significant difference in ChatGPT-3.5's performance across the subcategories of diagnosis, treatment, and others.

While ChatGPT-3.5 is not yet fully dependable for complex decision-making in medical oncology, it shows promise in the

field. In recent years, we have witnessed significant progress in neural networks, and the future of health care is becoming increasingly multimodal. Oncologists now rely on more than just text-based information when prescribing treatments. They consider a wide range of factors, including diverse image types, genomic data, and social determinants of health. However, in the past, developing multimodal machine learning models seemed like an overly ambitious goal. Thankfully, the landscape has changed, and we have seen exciting advancements in this area through various publications in 2022 and 2023 [29,30]. These studies have showcased the potential applications of multimodal models in the field of oncology, bringing us closer to a more comprehensive and holistic approach to cancer care.

Based on its performance in this study, we do not think that AI can aid oncologists in clinical decision-making at this time. However, it may excel in other tasks in the field [31]. Experts might look to language-generating AI to reduce the burden on humans who create questions and explanations for tests. However, it should be noted that ChatGPT-3.5 is not a useful tool without human supervision at this point, given its potential to fabricate references that may sound plausible but are incorrect [14,32,33]. Oncologists can also use it for administrative tasks such as drafting notes [34] or crafting communication messages for patients [11]. Additionally, while a previous study by Johnson et al [35] demonstrated that ChatGPT can be used by patients to answer common cancer myths and questions, the questions used in this study were already featured on the National Cancer Institute's webpage and were likely part of ChatGPT's training data [25] and fewer questions were used. We can infer from this study that the answers provided by ChatGPT still require review by an oncologist to ascertain their accuracy.

In the future, AI has the potential to assist oncologists in critical aspects such as determining optimal chemotherapy dosages [36] and aiding in diagnostics within fields like radiology and pathology [37]. By leveraging the capabilities of these advanced language models, health care professionals can access valuable insights and support in making informed decisions regarding

treatment plans. Moreover, patients can also reap the advantages of AI-driven technologies by receiving more accurate diagnoses and tailored treatment approaches, ultimately leading to improved outcomes and enhanced patient care [38].

This study does, however, have several important limitations. First, as ASCO-SEP only consists of MCQs, we did not challenge ChatGPT-3.5 with any other question formats (eg, open-ended), which may have yielded different results. Furthermore, MCQs may not fully reflect the complexity of clinical scenarios that oncologists face in their practice. Second, we did not test the variability of the answers provided by ChatGPT. Each question was presented to ChatGPT 3.5 only once, and the first answer was scored given that previous studies showed high consistency of ChatGPT answers [39]. Finally, we could have performed a qualitative assessment of ChatGPT-3.5 answers to gain insights into the etiology of its errors as a guide to future required improvements.

Conclusions

In conclusion, this study explored the capacity of ChatGPT-3.5's knowledge in medical oncology using the ASCO-SEP. We aimed to bridge the knowledge gaps surrounding the efficacy of AI-driven tools like ChatGPT-3.5 in supporting clinical decision-making. Our assessment revealed that while ChatGPT-3.5 shows promise for the future of AI in oncology, its current performance on ASCO-SEP underscores a pressing need for further refinement to meet the competency standards in this complex field.

Future evaluations of ChatGPT could extend to assessing its capability in clinical decision support, gauging its accuracy in real-life clinical scenarios, and its ease of integration into medical workflows. Evaluating GPT-4 as a resource to aid oncologists in clinical decision-making, an aspect not available during the tenure of this study, could significantly contribute to the field. The tool's facilitation of interdisciplinary collaboration among health care professionals and its impact on patient engagement and communication are other potential areas of investigation.

Conflicts of Interest

None declared.

References

1. Davenport T, Kalakota R. The potential for artificial intelligence in healthcare. *Future Healthc J* 2019;6(2):94-98 [FREE Full text] [doi: [10.7861/futurehosp.6-2-94](https://doi.org/10.7861/futurehosp.6-2-94)] [Medline: [31363513](https://pubmed.ncbi.nlm.nih.gov/31363513/)]
2. Brown T, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, et al. Language models are few-shot learners. *Adv Neural Inf Process Syst* 2020;33:1877-1901.
3. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. *PLOS Digit Health* 2023;2(2):e0000198 [FREE Full text] [doi: [10.1371/journal.pdig.0000198](https://doi.org/10.1371/journal.pdig.0000198)] [Medline: [36812645](https://pubmed.ncbi.nlm.nih.gov/36812645/)]
4. CHatGPT goes law school. University of Minnesota Law School. 2023. URL: <https://twin-cities.umn.edu/news-events/chatgpt-goes-law-school#:~:text=MINNEAPOLIS%2FST.achieved%20low%20but%20passing%20grades> [accessed 2023-12-08]
5. Khan RA, Jawaid M, Khan AR, Sajjad M. ChatGPT—reshaping medical education and clinical management. *Pak J Med Sci* 2023;39(2):605-607 [FREE Full text] [doi: [10.12669/pjms.39.2.7653](https://doi.org/10.12669/pjms.39.2.7653)] [Medline: [36950398](https://pubmed.ncbi.nlm.nih.gov/36950398/)]

6. King MR. A conversation on artificial intelligence, chatbots, and plagiarism in higher education. *Cell Mol Bioeng* 2023;16(1):1-2 [FREE Full text] [doi: [10.1007/s12195-022-00754-8](https://doi.org/10.1007/s12195-022-00754-8)] [Medline: [36660590](https://pubmed.ncbi.nlm.nih.gov/36660590/)]
7. Large language models. Wikipedia. 2023. URL: https://en.wikipedia.org/wiki/Large_language_model [accessed 2023-03-14]
8. Bajwa J, Munir U, Nori A, Williams B. Artificial intelligence in healthcare: transforming the practice of medicine. *Future Healthc J* 2021;8(2):e188-e194 [FREE Full text] [doi: [10.7861/fhj.2021-0095](https://doi.org/10.7861/fhj.2021-0095)] [Medline: [34286183](https://pubmed.ncbi.nlm.nih.gov/34286183/)]
9. Alfertshofer M, Hoch CC, Funk PF, Hollmann K, Wollenberg B, Knoedler S, et al. Sailing the seven seas: a multinational comparison of ChatGPT's performance on medical licensing examinations. *Ann Biomed Eng* 2023 [FREE Full text] [doi: [10.1007/s10439-023-03338-3](https://doi.org/10.1007/s10439-023-03338-3)] [Medline: [37553555](https://pubmed.ncbi.nlm.nih.gov/37553555/)]
10. Asch D. An interview with ChatGPT about health care. *NEJM Catal Innov Care Deliv* 2023;4(2) [FREE Full text]
11. Dave T, Athaluri SA, Singh S. ChatGPT in medicine: an overview of its applications, advantages, limitations, future prospects, and ethical considerations. *Front Artif Intell* 2023;6:1169595 [FREE Full text] [doi: [10.3389/frai.2023.1169595](https://doi.org/10.3389/frai.2023.1169595)] [Medline: [37215063](https://pubmed.ncbi.nlm.nih.gov/37215063/)]
12. Rich AS, Gureckis TM. Lessons for artificial intelligence from the study of natural stupidity. *Nat Mach Intell* 2019;1(4):174-180 [FREE Full text] [doi: [10.1038/s42256-019-0038-z](https://doi.org/10.1038/s42256-019-0038-z)]
13. Frey CB, Osborne MA. The future of employment: how susceptible are jobs to computerisation? *Technol Forecast Soc Change* 2017;114:254-280 [FREE Full text] [doi: [10.1016/j.techfore.2016.08.019](https://doi.org/10.1016/j.techfore.2016.08.019)]
14. Homolak J. Opportunities and risks of ChatGPT in medicine, science, and academic publishing: a modern promethean dilemma. *Croat Med J* 2023;64(1):1-3 [FREE Full text] [doi: [10.3325/cmj.2023.64.1](https://doi.org/10.3325/cmj.2023.64.1)] [Medline: [36864812](https://pubmed.ncbi.nlm.nih.gov/36864812/)]
15. Melnikov AA, Nautrup HP, Krenn M, Dunjko V, Tiersch M, Zeilinger A, et al. Active learning machine learns to create new quantum experiments. *Proc Natl Acad Sci U S A* 2018;115(6):1221-1226 [FREE Full text] [doi: [10.1073/pnas.1714936115](https://doi.org/10.1073/pnas.1714936115)] [Medline: [29348200](https://pubmed.ncbi.nlm.nih.gov/29348200/)]
16. van Dis EAM, Bollen J, Zuidema W, van Rooij R, Bockting CL. ChatGPT: five priorities for research. *Nature* 2023;614(7947):224-226. [doi: [10.1038/d41586-023-00288-7](https://doi.org/10.1038/d41586-023-00288-7)] [Medline: [36737653](https://pubmed.ncbi.nlm.nih.gov/36737653/)]
17. ASCO-SEP®. American Society of Clinical Oncology® Store. 2023. URL: <https://shop.asco.org/asco-sep/#:~:text=ASCO%2DSEP%2%AE%20is%20a,and%20cancer%20in%20elderly%20patients> [accessed 2023-03-15]
18. ASCO-SEP 6th edition self-evaluation. American Society of Clinical Oncology® Education. 2023. URL: <https://education.asco.org/product-details/asco-sep-6th-edition-self-evaluation> [accessed 2023-05-05]
19. Gilson A, Safranek CW, Huang T, Socrates V, Chi L, Taylor RA, et al. How does ChatGPT perform on the United States medical licensing examination? The implications of large language models for medical education and knowledge assessment. *JMIR Med Educ* 2023;9:e45312 [FREE Full text] [doi: [10.2196/45312](https://doi.org/10.2196/45312)] [Medline: [36753318](https://pubmed.ncbi.nlm.nih.gov/36753318/)]
20. Explanations for the 2020-2022 official step 2 CK practice questions. Ben White. 2023. URL: <https://www.benwhite.com/medicine/explanations-for-the-2020-2021-official-step-2-ck-practice-questions/> [accessed 2023-10-22]
21. Das D, Kumar N, Longjam LA, Sinha R, Roy AD, Mondal H, et al. Assessing the capability of ChatGPT in answering first- and second-order knowledge questions on microbiology as per competency-based medical education curriculum. *Cureus* 2023;15(3):e36034 [FREE Full text] [doi: [10.7759/cureus.36034](https://doi.org/10.7759/cureus.36034)] [Medline: [37056538](https://pubmed.ncbi.nlm.nih.gov/37056538/)]
22. Sinha RK, Roy AD, Kumar N, Mondal H. Applicability of ChatGPT in assisting to solve higher order problems in pathology. *Cureus* 2023;15(2):e35237 [FREE Full text] [doi: [10.7759/cureus.35237](https://doi.org/10.7759/cureus.35237)] [Medline: [36968864](https://pubmed.ncbi.nlm.nih.gov/36968864/)]
23. Chan MW, Eppich WJ. The keyword effect: a grounded theory study exploring the role of keywords in clinical communication. *AEM Educ Train* 2020;4(4):403-410 [FREE Full text] [doi: [10.1002/aet2.10424](https://doi.org/10.1002/aet2.10424)] [Medline: [33150283](https://pubmed.ncbi.nlm.nih.gov/33150283/)]
24. Keywords/Buzzwords on step. Reddit. 2023. URL: https://www.reddit.com/r/step1/comments/o56ho3/keywordsbuzzwords_on_step/?rdt=55189 [accessed 2023-10-22]
25. Schade M. How ChatGPT and our language models are developed. OpenAI. 2023. URL: <https://help.openai.com/en/articles/7842364-how-chatgpt-and-our-language-models-are-developed> [accessed 2023-10-27]
26. Takagi S, Watari T, Erabi A, Sakaguchi K. Performance of GPT-3.5 and GPT-4 on the Japanese medical licensing examination: comparison study. *JMIR Med Educ* 2023;9:e48002 [FREE Full text] [doi: [10.2196/48002](https://doi.org/10.2196/48002)] [Medline: [37384388](https://pubmed.ncbi.nlm.nih.gov/37384388/)]
27. Kumah-Crystal Y, Mankowitz S, Embi P, Lehmann CU. ChatGPT and the clinical informatics board examination: the end of unproctored maintenance of certification? *J Am Med Inform Assoc* 2023;30(9):1558-1560. [doi: [10.1093/jamia/ocad104](https://doi.org/10.1093/jamia/ocad104)] [Medline: [37335851](https://pubmed.ncbi.nlm.nih.gov/37335851/)]
28. Arif TB, Munaf U, Ul-Haque I. The future of medical education and research: Is ChatGPT a blessing or blight in disguise? *Med Educ Online* 2023;28(1):2181052 [FREE Full text] [doi: [10.1080/10872981.2023.2181052](https://doi.org/10.1080/10872981.2023.2181052)] [Medline: [36809073](https://pubmed.ncbi.nlm.nih.gov/36809073/)]
29. Boehm KM, Khosravi P, Vanguri R, Gao J, Shah SP. Harnessing multimodal data integration to advance precision oncology. *Nat Rev Cancer* 2022;22(2):114-126 [FREE Full text] [doi: [10.1038/s41568-021-00408-3](https://doi.org/10.1038/s41568-021-00408-3)] [Medline: [34663944](https://pubmed.ncbi.nlm.nih.gov/34663944/)]
30. Foersch S, Glasner C, Woerl AC, Eckstein M, Wagner DC, Schulz S, et al. Multistain deep learning for prediction of prognosis and therapy response in colorectal cancer. *Nat Med* 2023;29(2):430-439 [FREE Full text] [doi: [10.1038/s41591-022-02134-1](https://doi.org/10.1038/s41591-022-02134-1)] [Medline: [36624314](https://pubmed.ncbi.nlm.nih.gov/36624314/)]
31. Liu J, Wang C, Liu S. Utility of ChatGPT in clinical practice. *J Med Internet Res* 2023;25:e48568 [FREE Full text] [doi: [10.2196/48568](https://doi.org/10.2196/48568)] [Medline: [37379067](https://pubmed.ncbi.nlm.nih.gov/37379067/)]
32. David Smerdon. X. 2023. URL: <https://twitter.com/dsmerdon/status/1618816703923912704> [accessed 2023-04-30]
33. Teresa Kubacka. X. 2023. URL: https://twitter.com/paniterka_ch/status/1599893718214901760 [accessed 2023-04-30]

34. Patel SB, Lam K. ChatGPT: the future of discharge summaries? *Lancet Digit Health* 2023;5(3):e107-e108 [FREE Full text] [doi: [10.1016/S2589-7500\(23\)00021-3](https://doi.org/10.1016/S2589-7500(23)00021-3)] [Medline: [36754724](https://pubmed.ncbi.nlm.nih.gov/36754724/)]
35. Johnson SB, King AJ, Warner EL, Aneja S, Kann BH, Bylund CL. Using ChatGPT to evaluate cancer myths and misconceptions: artificial intelligence and cancer information. *JNCI Cancer Spectr* 2023;7(2):pkad015 [FREE Full text] [doi: [10.1093/jncics/pkad015](https://doi.org/10.1093/jncics/pkad015)] [Medline: [36929393](https://pubmed.ncbi.nlm.nih.gov/36929393/)]
36. Londhe VY, Bhasin B. Artificial intelligence and its potential in oncology. *Drug Discov Today* 2019;24(1):228-232 [FREE Full text] [doi: [10.1016/j.drudis.2018.10.005](https://doi.org/10.1016/j.drudis.2018.10.005)] [Medline: [30342246](https://pubmed.ncbi.nlm.nih.gov/30342246/)]
37. Luchini C, Pea A, Scarpa A. Artificial intelligence in oncology: current applications and future perspectives. *Br J Cancer* 2022;126(1):4-9 [FREE Full text] [doi: [10.1038/s41416-021-01633-1](https://doi.org/10.1038/s41416-021-01633-1)] [Medline: [34837074](https://pubmed.ncbi.nlm.nih.gov/34837074/)]
38. Huang S, Cai N, Pacheco PP, Narrandes S, Wang Y, Xu W. Applications of Support Vector Machine (SVM) learning in cancer genomics. *Cancer Genomics Proteomics* 2018;15(1):41-51 [FREE Full text] [doi: [10.21873/cgp.20063](https://doi.org/10.21873/cgp.20063)] [Medline: [29275361](https://pubmed.ncbi.nlm.nih.gov/29275361/)]
39. Suárez A, García VDF, Algar J, Gómez Sánchez M, de Pedro ML, Freire Y. Unveiling the ChatGPT phenomenon: evaluating the consistency and accuracy of endodontic question answers. *Int Endod J* 2023:108-113 [FREE Full text] [doi: [10.1111/iej.13985](https://doi.org/10.1111/iej.13985)] [Medline: [37814369](https://pubmed.ncbi.nlm.nih.gov/37814369/)]

Abbreviations

AI: artificial intelligence

ASCO-SEP: American Society of Clinical Oncology Self-Evaluation Program

MCQ: multiple choice question

USMLE: United States Medical Licensing Examination

Edited by K El Emam; submitted 02.07.23; peer-reviewed by D Hu, S Matsuda, J Luo; comments to author 28.07.23; revised version received 05.10.23; accepted 19.11.23; published 12.01.24.

Please cite as:

*Odabashian R, Bastin D, Jones G, Manzoor M, Tangestaniapour S, Assad M, Lakhani S, Odabashian M, McGee S
Assessment of ChatGPT-3.5's Knowledge in Oncology: Comparative Study with ASCO-SEP Benchmarks
JMIR AI 2024;3:e50442*

URL: <https://ai.jmir.org/2024/1/e50442>

doi: [10.2196/50442](https://doi.org/10.2196/50442)

PMID:

©Roupen Odabashian, Donald Bastin, Georden Jones, Maria Manzoor, Sina Tangestaniapour, Malke Assad, Sunita Lakhani, Maritsa Odabashian, Sharon McGee. Originally published in JMIR AI (<https://ai.jmir.org>), 12.01.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR AI, is properly cited. The complete bibliographic information, a link to the original publication on <https://www.ai.jmir.org/>, as well as this copyright and license information must be included.

Original Paper

The Impact of Expectation Management and Model Transparency on Radiologists' Trust and Utilization of AI Recommendations for Lung Nodule Assessment on Computed Tomography: Simulated Use Study

Lotte J S Ewals¹, MSc; Lynn J J Heesterbeek², MSc; Bin Yu³, PhD; Kasper van der Wulp¹, MD; Dimitrios Mavroeidis⁴, PhD; Mathias Funk⁵, PhD; Chris C P Snijders⁶, Prof Dr; Igor Jacobs⁷, PhD; Joost Nederend¹, MD, PhD; Jon R Pluyter², PhD; e/MTIC Oncology group⁸

¹Catharina Cancer Institute, Catharina Hospital Eindhoven, Eindhoven, Netherlands

²Department of Experience Design, Royal Philips, Eindhoven, Netherlands

³Research Center for Marketing and Supply Chain Management, Nyenrode Business University, Breukelen, Netherlands

⁴Department of Data Science, Philips Research, Eindhoven, Netherlands

⁵Department of Industrial Design, Eindhoven University of Technology, Eindhoven, Netherlands

⁶Department of Human Technology Interaction, Eindhoven University of Technology, Eindhoven, Netherlands

⁷Department of Hospital Services and Informatics, Philips Research, Eindhoven, Netherlands

⁸See acknowledgments, Eindhoven, Netherlands

Corresponding Author:

Lotte J S Ewals, MSc

Catharina Cancer Institute, Catharina Hospital Eindhoven

Michelangelolaan 2

Eindhoven, 5623 EJ

Netherlands

Phone: 31 40 239 9111

Email: lotte.ewals@catharinaziekenhuis.nl

Abstract

Background: Many promising artificial intelligence (AI) and computer-aided detection and diagnosis systems have been developed, but few have been successfully integrated into clinical practice. This is partially owing to a lack of user-centered design of AI-based computer-aided detection or diagnosis (AI-CAD) systems.

Objective: We aimed to assess the impact of different onboarding tutorials and levels of AI model explainability on radiologists' trust in AI and the use of AI recommendations in lung nodule assessment on computed tomography (CT) scans.

Methods: In total, 20 radiologists from 7 Dutch medical centers performed lung nodule assessment on CT scans under different conditions in a simulated use study as part of a 2×2 repeated-measures quasi-experimental design. Two types of AI onboarding tutorials (reflective vs informative) and 2 levels of AI output (black box vs explainable) were designed. The radiologists first received an onboarding tutorial that was either informative or reflective. Subsequently, each radiologist assessed 7 CT scans, first without AI recommendations. AI recommendations were shown to the radiologist, and they could adjust their initial assessment. Half of the participants received the recommendations via black box AI output and half received explainable AI output. Mental model and psychological trust were measured before onboarding, after onboarding, and after assessing the 7 CT scans. We recorded whether radiologists changed their assessment on found nodules, malignancy prediction, and follow-up advice for each CT assessment. In addition, we analyzed whether radiologists' trust in their assessments had changed based on the AI recommendations.

Results: Both variations of onboarding tutorials resulted in a significantly improved mental model of the AI-CAD system (informative $P=.01$ and reflective $P=.01$). After using AI-CAD, psychological trust significantly decreased for the group with explainable AI output ($P=.02$). On the basis of the AI recommendations, radiologists changed the number of reported nodules in 27 of 140 assessments, malignancy prediction in 32 of 140 assessments, and follow-up advice in 12 of 140 assessments. The changes were mostly an increased number of reported nodules, a higher estimated probability of malignancy, and earlier follow-up.

The radiologists' confidence in their found nodules changed in 82 of 140 assessments, in their estimated probability of malignancy in 50 of 140 assessments, and in their follow-up advice in 28 of 140 assessments. These changes were predominantly increases in confidence. The number of changed assessments and radiologists' confidence did not significantly differ between the groups that received different onboarding tutorials and AI outputs.

Conclusions: Onboarding tutorials help radiologists gain a better understanding of AI-CAD and facilitate the formation of a correct mental model. If AI explanations do not consistently substantiate the probability of malignancy across patient cases, radiologists' trust in the AI-CAD system can be impaired. Radiologists' confidence in their assessments was improved by using the AI recommendations.

(JMIR AI 2024;3:e52211) doi:[10.2196/52211](https://doi.org/10.2196/52211)

KEYWORDS

application; artificial intelligence; AI; computer-aided detection or diagnosis; CAD; design; human centered; human computer interaction; HCI; interaction; mental model; radiologists; trust

Introduction

Background

Lung cancer is one of the leading causes of cancer-related deaths worldwide [1]. Early detection of lung cancer is essential to provide curative treatment and improve survival. However, detecting and diagnosing lung cancer using computed tomography (CT) scans can be challenging. On CT scans, early lung cancer can be seen as a small nodule. However, these nodules can also be benign. The risk of malignancy depends on various patient factors and lung nodule features, such as the morphology, size, and number of lung nodules. Nodules that are challenging to detect can, for instance, be small, and their perceptibility might be hampered by their location close to normal lung tissue that is visually similar on a CT scan, such as blood vessels or bronchi [2-5]. As a result, radiologists may overlook or misdiagnose lung nodules on CT scans. A previous study showed that radiologists missed 15% of all lung cancer cases on screening CT scans. Of these missed cancers diagnoses, 35% were not visible on the scan, 50% were not detected by the radiologist, and 15% were detected but not diagnosed as cancer [6].

A recent approach to improve the detection and diagnosis of lung nodules on CT scans is the use of artificial intelligence (AI) models. Diagnostic assistance from AI models that provide recommendations for radiologists is referred to as AI-based computer-aided detection or diagnosis (AI-CAD) [7]. Many studies have been published on AI models for assessing lung nodules on CT scans, showing promising performance with sensitivities for detection of up to 98.1% and a mean of only 2 false-positives (FPs) per scan [8,9].

Although many AI models and AI-CAD systems have been developed, few are used in clinical practice. Although most studies on AI for lung nodule assessment focus on the development and stand-alone performance of AI models [8,10,11], few studies have focused on user interaction with AI models in the clinical context beyond the theoretical level [12-16]. However, human-AI interaction is essential to enable radiologists to comprehend and effectively use AI recommendations in their tasks, ultimately achieving the highest levels of diagnostic quality and efficiency.

Trust is of great importance in the interactions and collaborations between radiologists and AI-CAD systems [15,17-20]. Trust influences the end users' level of reliance on AI recommendations, and hence, it influences the performance of AI-assisted end users [18,19]. If the user has very little trust in the system, the potential benefits of AI-CAD will be reduced because of disuse, whereas too much trust in the system leads to overreliance and can result in mistakes that would not have been made without using the AI-CAD system [15,18].

Trust is a dynamic process. Trust changes over time and across situations and is influenced by many factors. For example, trust varies based on the reliability of the AI system, the design of the system, the personal characteristics of the user, prior interactions and experience, and moderating factors such as workload and sociocultural context [18,21-25]. Some of these factors can be influenced through the design of the system, with the aim of achieving the formation of appropriate trust. Trust calibration refers to interventions that facilitate the formation of an appropriate trust level by aligning a person's trust in the AI with the capabilities of the AI [26,27]. In this study, we introduced 2 instruments aimed at appropriate trust calibration at different time points of use. First, an onboarding tutorial aimed to set the right expectations before initial use. Second, AI model explainability as an information cue available to clinical users during use to judge the credibility of the arguments underpinning the AI model prediction.

We aimed to assess whether radiologists' trust in AI-CAD systems and their use of AI recommendations in lung nodule assessments on CT scans were affected by different onboarding tutorials and by different levels of AI model explainability.

Theoretical Argumentation

Trust Definitions

Different definitions and measures exist for trust [15]. In this study, we considered trust from 2 complementary perspectives, a cognitive perspective and a behavioral perspective [23].

From the cognitive perspective, we explored the users' mental model and psychological trust. The *mental model* represents a person's "static knowledge about the system: its significant features, how it functions, how different components affect others, and how its components will behave when confronted with various factors and influences" [24]. In short, the mental

model is the user's understanding of the AI system. A correct mental model is expected to contribute to appropriate trust calibration between the user's trust in an AI system and the trustworthiness of the system [25]. User's *psychological trust* refers to "the extent to which a user is confident in, and willing to act on the basis of, the recommendations, actions, and decisions of an artificially intelligent decision aid" [28]. Because radiologists gain experience and learn through the process of assessing CT cases with the AI-CAD tool and actually see what the system is capable of, they are expected to have an improved mental model of (hypothesis 1a) and psychological trust in (hypothesis 1b) the AI-CAD system after using the AI-CAD system compared with before using the system.

However, holding a positive attitude toward the AI-CAD system does not mean that the user will also act in line with its recommendations. Therefore, we also adopted a behavioral lens by examining whether trust was reflected in the *use of the AI recommendations* (reliance and compliance) and the corresponding impact on decision outcomes [29,30]. The decision of whether radiologists use AI recommendations depends not only on their overall trust in the AI-CAD system but also on their agreement with the specific AI recommendations for a given case. As the AI recommendations function as a second reader, it is expected that radiologists' confidence in their assessments will be higher when they are assisted by AI-CAD than without assistance (hypothesis 2).

Onboarding Tutorials

Research on how to ensure that radiologists have appropriate expectations of the system's capabilities and limitations is limited [27]. As suggested by Cai et al [31], when clinical practitioners are first introduced to an AI system, a human-AI onboarding process can be crucial for them to determine how they will partner with AI in practice. Therefore, an onboarding tutorial to inform radiologists about the capabilities and limitations of the AI-CAD system is expected to improve radiologists' mental model of (hypothesis 3a) and psychological trust in the AI-CAD system (hypothesis 3b).

Moreover, critical reflection on one's experience is essential for developing competence and self-awareness [32]. Hence, it is hypothesized that critical reflection and feedback built through a reflective onboarding tutorial will lead to a more improved mental model of (hypothesis 4a) and psychological trust in (hypothesis 4b) the AI-CAD system than an informative onboarding tutorial. Furthermore, it is expected to be easier for radiologists to understand whether an AI suggestion should be followed because of their understanding of the AI-CAD system from reflective onboarding, especially when they are not fully sure of their own assessment. Therefore, it is expected that radiologists who receive reflective onboarding will use the AI recommendations more often than radiologists who receive informative onboarding (hypothesis 5).

Levels of AI Model Explainability

In addition, radiologists are expected to better judge whether they can trust an AI recommendation when the AI model discloses the reasoning behind its recommendations (explainable AI models) compared with black box models. Hence, it is hypothesized that after using the AI-CAD system, radiologists assisted with explainable AI output have an improved mental model of (hypothesis 6a) and psychological trust in (hypothesis 6b) the AI-CAD system than radiologists assisted with black box AI output. Because radiologists can see the reasoning behind the recommendations when receiving explainable AI output, it is expected that they will use the AI recommendations more often than radiologists assisted with black box AI output (hypothesis 7).

Methods

Overview

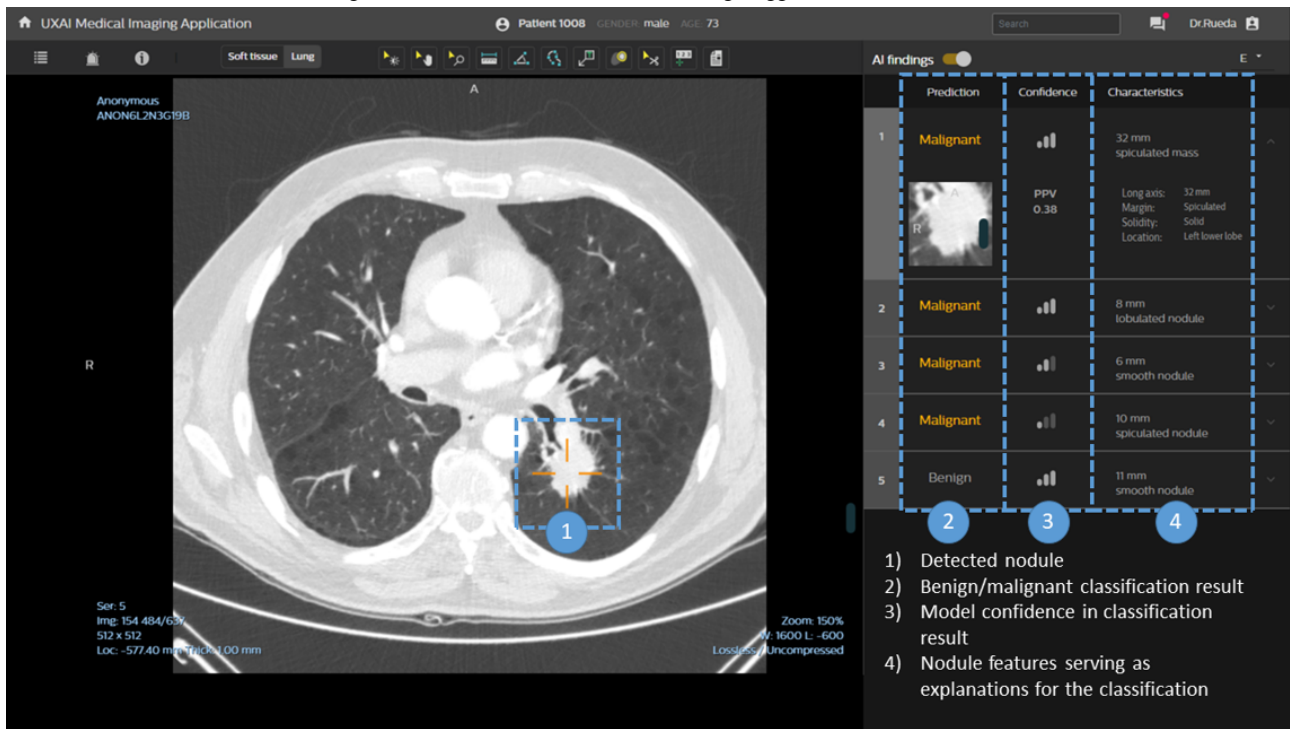
We tested the hypotheses using a 2×2 repeated-measures quasi-experimental design: informative versus reflective onboarding tutorial and black box versus explainable AI output. In this simulated use study, we aimed to realistically mimic clinical practice [33,34]. Realistic clinical simulations allow participants to engage with the setup in real-world clinical scenarios and encourage participants to authentically execute the study as if they are performing their clinical work.

Prototype

Image Viewer

A medical image-viewing prototype was developed to enable radiologists to assess incidental lung nodules on cardiac CT scans with and without the assistance of an AI-CAD system. The AI recommendations were implemented as a second reader, allowing the radiologist to first assess the cases independently. The interface was designed based on the literature, brainstorming, and feedback sessions with radiologists and design specialists and was iteratively optimized for the 2 variations of onboarding tutorials (reflective vs informative) and 2 variations of AI outputs (black box vs explainable). The final user interface is shown in [Figure 1](#). We aimed to realistically simulate the radiologists' clinical setup to facilitate proper engagement of the participants with the task of lung nodule assessment. The user setup was designed to simulate clinical practice as realistically as possible. The developed interface was shown to the radiologists on a monitor, which was placed in a separate silent room. This room was inside the hospital, and lights could be dimmed if the radiologists preferred it, comparable with their own working space. Similar to the picture archiving and communication system used in clinical practice to assess CT scans, radiologists could scroll through the images, zoom in, measure, and change the windowing level between the soft tissue and lung setting using a computer mouse.

Figure 1. Medical image-viewing prototype in the explainable artificial intelligence (AI) condition. In the black box AI condition, users could not see the nodule characteristics column on the right side of the screen. When the AI findings toggle is off, all AI recommendations will be hidden to the users.



Clinical Data

To further increase study engagement and realism, the use scenarios were based on real-world patient cases. We retrospectively selected 10 CT angiography scans with incidental pulmonary nodules from a large Dutch clinical hospital. Scans acquired between 2008 and 2015 were used because the 5-year outcomes of these patients are known: whether they developed lung cancer. An expert radiologist selected the cases for this study. Of the 10 selected scans, we used 3 for onboarding and 7 for testing the impact of the design interventions. All CT scans were performed on patients with lung cancer. By selecting the 7 CT cases, we aimed to obtain a diverse mix of assessment complexity by including both lower and higher suspicious nodules (based on size, spiculation, and solidity) and nodules at easier and more difficult locations (such as against the veins or pleura). The characteristics of the 7 CT cases and the findings of the AI model for these cases are presented in [Multimedia Appendix 1](#).

AI Model

To detect and estimate the malignancy of lung nodules on the CT scans, the pretrained AI framework developed by Trajanovski et al [35] was applied. This framework relies on a 2-stage process, where the first stage performs nodule detection and the second stage assigns a malignancy probability to the detected nodules. Among the validated nodule detectors, the best performance was achieved by the nodule detector developed by Liao et al [36]. This nodule detector is based on deep learning models, more precisely, convolutional neural networks. The nodules detected by the nodule detector are provided as input to the second stage of the framework that assigns the cancer malignancy probabilities. The second stage of the framework is based on a convolutional neural network that was trained

using the publicly available National Lung Screening Trial data set [37].

During inference, the model takes a CT scan as input and automatically produces a list of nodule locations (x,y,z), their radii, and malignancy probabilities. The prototype, described previously, ensures that this information is displayed intuitively to the clinicians. The article by Liao et al [36] provides all the relevant details regarding the training process and performance validation.

In this study, the AI model proposed by Trajanovski et al [35] was used without any additional fine-tuning. Specifically, the model weights remained unchanged. The sole adjustment involved calibrating (or rescaling) the output of the model to accommodate the changed distribution of malignant cases ([Multimedia Appendix 2](#) [35,38,39]).

AI Recommendations

The AI model recommendations were provided using 4 information cues ([Multimedia Appendix 3](#)):

1. Detected nodules (shown by target mark directly on the CT scan)
2. Benign or malignant classification per nodule (malignant nodules are highlighted in orange color)
3. Model confidence in the benign or malignant classification (shown as the negative predictive value [NPV] or positive predictive value [PPV] score and an intuitive icon representing high, medium, or low confidence)
4. In the explainable AI output variant: nodule features serving as an explanation for the classification

AI nodule detection and benign or malignant classification (cues 1 and 2) were obtained using the described AI model [35]. The number of lung nodules detected by the AI model varied

between 1 and 5 per scan. The AI model found at least one true-positive lung nodule in each case and found one or more FP nodules in 4 of 7 cases. For more information about the AI findings, see Table S1 in [Multimedia Appendix 1](#).

Confidence in the malignancy classification (cue 3) was given by means of PPVs for malignant predictions, indicating the probability that nodules with malignant predictions were actually malignant, and by means of NPVs for benign predictions, indicating the probability that nodules with a benign prediction were actually benign. The PPV was 0.25 (low confidence), 0.30 (medium confidence), or 0.38 (high confidence), and the NPV was 0.94 (low confidence), 0.97 (medium confidence), or >0.99 (high confidence; for an explanation of how the PPVs and NPVs were calculated, see [Multimedia Appendix 2](#)). In addition, confidence was shown by means of a small bar graph, indicating low, medium, or high model confidence.

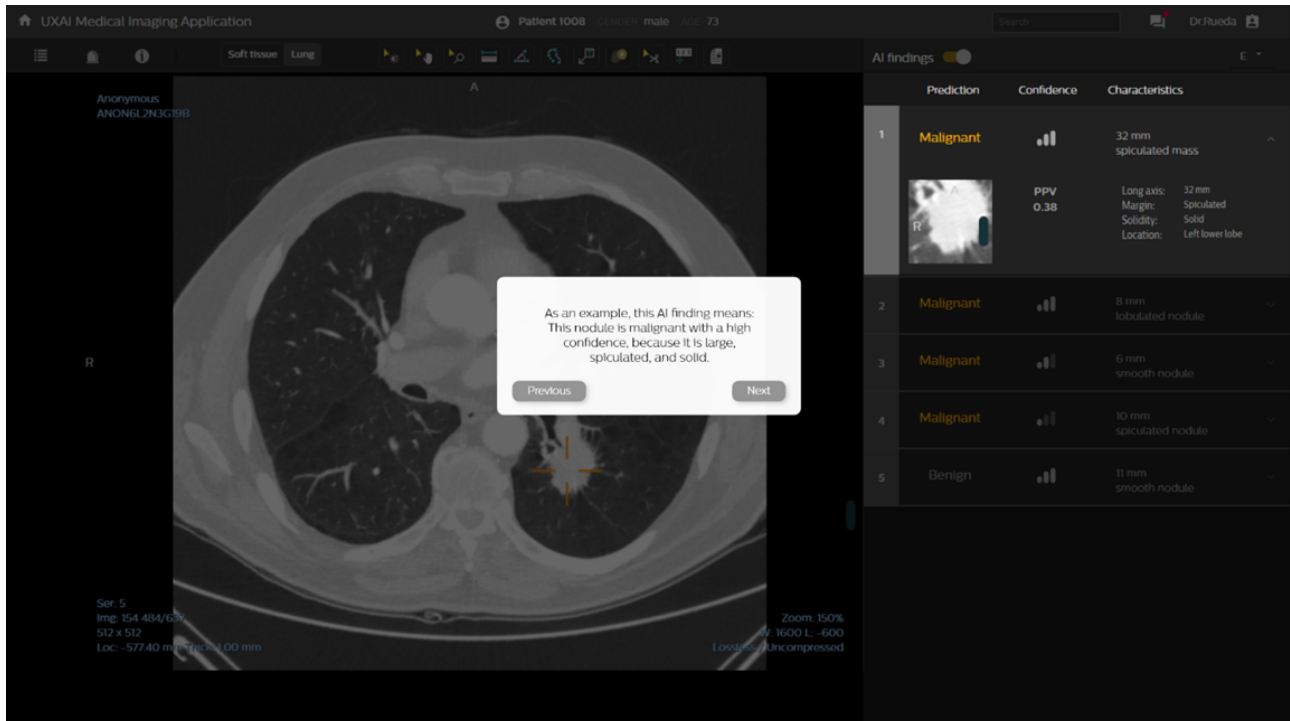
Two levels of AI transparency were tested: black box AI output and explainable AI output. Black box output indicates that radiologists did not see what the malignancy estimation was based on. The explainable AI output variant provided the same information as the black box AI output variant and additionally showed the characteristics of the lung nodules (cue 4); this information was expected to help in understanding and interpreting the predictions of the AI-CAD system ([Figure 1](#), right column). For each lung nodule, the following lung nodule

characteristics were provided: long axis diameter, solidity, margin characteristics, and location. The nodule characteristics were not provided by the AI model and were therefore realistically simulated, which is in agreement with related research [40] via manual annotation by 2 expert radiologists in consensus. However, the participants were not aware of the simulation; therefore, from the radiologists' perspective, the characteristics were AI generated as well [41]. For an overview of the information cues for the AI recommendations, see [Multimedia Appendix 3](#).

Onboarding Tutorials

Two variations of onboarding tutorials were designed: informative onboarding and reflective onboarding. During informative onboarding, radiologists passively received a stepwise introduction of the AI capabilities and common pitfalls so that they could acquire a realistic mental model of the system ([Figure 2](#)). The AI model's capabilities and pitfalls were illustrated in the onboarding tutorial with 3 CT scans that showed obvious cancer cases, FP nodules, and false-negative nodules. For an overview of all implemented questions and explanations, see [Multimedia Appendix 3](#). During reflective onboarding, radiologists additionally engaged in active reflection. They received cognitive feedback on 4 questions that they had to answer to check whether their mental model of the AI-CAD system was correct.

Figure 2. Onboarding tutorial in the informative onboarding condition, which provided a stepwise introduction of artificial intelligence (AI) capabilities and limitations using example patient cases. In the reflective onboarding condition, an additional question-answer dialog was triggered to provide feedback on whether the user's expectations of the AI capabilities and limitations were correct.



Study Protocol

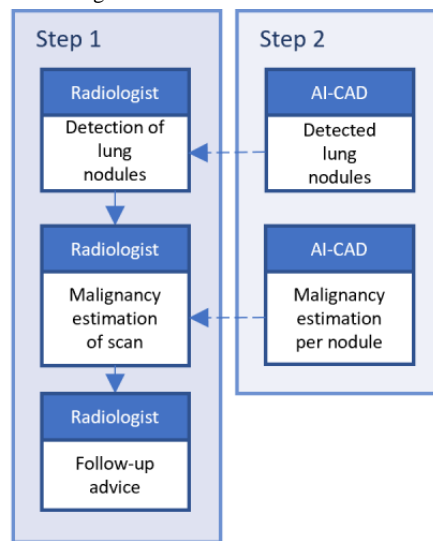
For this study, physicians were eligible for participation if they were radiologists, nuclear radiologists, or radiology residents. We will refer to the participants as *radiologists*. Several effects were to be tested; we used a power of 80%. For the mental

model differences between radiologists, we based our sample size calculation on a comparison of means of 2 versus 3 (SD 0.5). This led to a necessary sample size of 12 radiologists. For the psychological trust differences, we based the sample size calculation on a comparison of means of 0.5 versus 0.75 (SD 0.1). This resulted in a sample size of at least 8 radiologists.

The differences in the use of AI recommendations were based on a comparison of proportions in the order of magnitude of 30% versus 10%. This leads to a necessary sample size of 124 comparisons if we assume that the intraclass coefficient is low. Eventually, 20 radiologists were included in this study, all of whom assessed 7 CT scans for a total of 140 recommendations [42]. In this 2x2 repeated-measures design, the radiologists were divided into 4 groups, each of which consisted of 5 radiologists. After onboarding in one of the 2 conditions, using 3 CT scans, each radiologist assessed the 7 CT scans. In addition to the CT scans, each patient’s age and gender were provided

because radiologists also use the patient context when they assess CT scans in clinical practice. First, the radiologists assessed the scans without observing the AI output. They reported the nodules they detected, estimated the malignancy probability for the patient case (not per nodule, unlike the AI model), and provided follow-up advice. Subsequently, the AI recommendations were presented, and the radiologists could adjust their initial assessments. The nodules detected by AI and the AI malignancy estimations might trigger the radiologists to change their initial assessments. This process is visualized in the flow diagram in Figure 3.

Figure 3. Flow diagram showing the clinical decisions of radiologists, which might potentially be influenced by the outcomes of the artificial intelligence model. The detected nodules may influence the malignancy estimation, and the malignancy estimation may influence the follow-up advice. AI-CAD: artificial intelligence–based computer-aided detection or diagnosis.

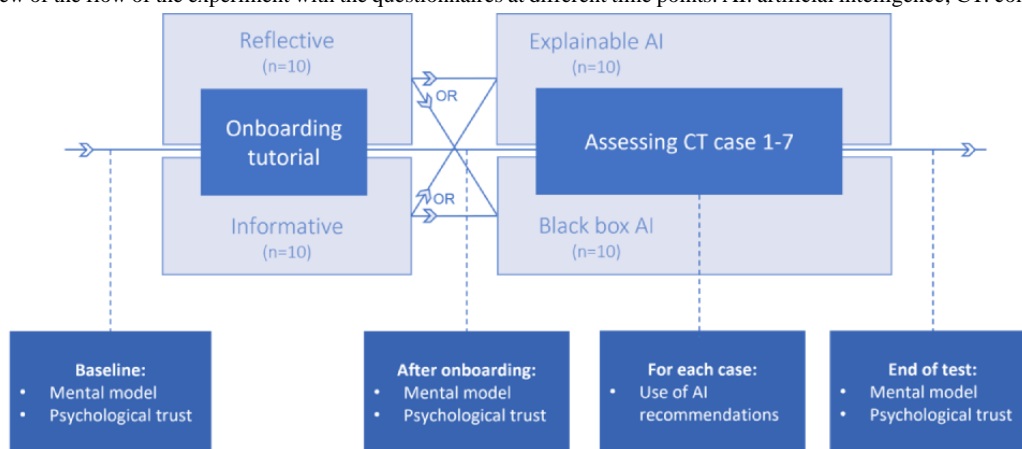


Measures for Trust

To evaluate the effects of the 2 types of AI onboarding tutorials and the 2 levels of explainability of AI outputs on radiologists’ trust in AI and their use of AI recommendations, participants were requested to complete questionnaires on 3 aspects: the

radiologists’ mental model of the AI-CAD system’s capabilities and pitfalls, psychological trust in the AI-CAD system, and the use of AI recommendations. These questionnaires were completed at different time points, as schematically shown in Figure 4.

Figure 4. Overview of the flow of the experiment with the questionnaires at different time points. AI: artificial intelligence; CT: computed tomography.



Mental Model

The mental model questionnaire measured the radiologists’ understanding of the AI capabilities and limitations to uncover whether their expectations of the AI-CAD system were appropriate. Of the 11 questions in this questionnaire, 5

questions were related to nodule detection and 6 were related to malignancy prediction (see the full questionnaire in Multimedia Appendix 4). Questions could be answered with *yes*, *no*, or *I do not know*. Depending on whether the assessment was correct as compared with the true AI capabilities, a score

of 1 (correct) or 0 (incorrect or *I do not know*) was assigned per question, resulting in summed scores between 0 and 11. A higher score implies a better understanding of the AI capabilities. The mental model was measured before onboarding, after onboarding, and after assessing the 7 CT scans.

Psychological Trust

To measure the radiologists' psychological trust in the AI-CAD system, a questionnaire was derived from the study by Ashoori and Weisz [43] and adapted to fit this study (see the full questionnaire in [Multimedia Appendix 4](#)). This questionnaire examined overall trustworthiness, reliability, technical competence, and personal attachment. An example of a statement is "This model is trustworthy." The 12 statements about the AI model had to be answered with a score between 1 (strongly disagree) and 5 (strongly agree). For the negatively phrased questions, scores were reversed for the data analysis so that for all questions, a higher score reflected more trust in the AI-CAD system. Subsequently, the scores for the 12 questions were averaged. The psychological trust of each participant was measured before onboarding, after onboarding, and after assessing the 7 CT scans.

Use of AI Recommendations

To evaluate the radiologists' use of the AI recommendations, their assessments and confidence in their assessments—first without and then with AI assistance—were recorded in a questionnaire. AI recommendation use was measured at 3 assessment levels: number of detected nodules, malignancy probability, and follow-up advice. Therefore, the questionnaire included questions about the number of found nodules, the malignancy probability (at the patient level) as a percentage, and the follow-up advice according to the Fleischner guidelines [44]. The follow-up advice had to be scored with a score of 1 (consider CT at 3 months, positron emission tomography-CT, or tissue sampling), 2 (CT at 3-6 months), 3 (CT at 6-12 months), 4 (CT at 12 months), or 5 (no routine follow-up). A lower score indicated earlier follow-up. In addition, the confidence of the given answers at each assessment level had to be rated with a score between 1 (not confident at all) and 5 (very confident). The complete questionnaire is provided in [Multimedia Appendix 4](#). Participants were requested to complete this questionnaire while assessing without AI assistance and with AI assistance for each CT case.

Analyses

Mental Model and Psychological Trust

Changes in the mental model and psychological trust were assessed by comparing the scores before and after onboarding, and the scores after onboarding and at the end of the test, that is, after assessing all 7 CT scans. These changes were assessed for all radiologists together, for the 2 onboarding tutorial groups separately, and for the 2 AI output groups separately. The changes in scores were compared between the 2 onboarding tutorial groups and between the 2 AI output groups to analyze whether the types of onboarding tutorials and level of AI explainability influenced radiologists' initial trust and maintenance of trust during CT assessment. In addition, we analyzed whether the changes in mental model and

psychological trust scores were influenced by any of the following characteristics of the radiologists: age, gender, years of experience, how often they assessed lungs on CT as part of their job, how eager they were to try new information technologies, and how frequently they used AI-CAD tools.

Use of AI Recommendations

The use of AI recommendations was assessed by analyzing the number of cases in which radiologists adjusted the number of found nodules, the malignancy probability, and the follow-up advice after viewing the AI-CAD recommendations. In addition, we analyzed whether the radiologist's confidence in the assessments of the number of nodules, the malignancy prediction, and the follow-up advice changed after viewing the AI recommendations and whether their confidence increased or decreased. The use of AI recommendations and the impact on radiologists' confidence were compared between the groups of onboarding tutorials and between the groups of AI output.

Secondary Analyses

Additional Analyses and Use of AI Recommendations

In addition, the impact of agreeing or disagreeing with the AI detected nodules was evaluated. We analyzed whether the use of AI recommendations and radiologists' confidence in their assessments were affected by 2 factors: first, whether the same or different nodules were found by the AI as compared with the radiologist and, second, whether the radiologist changed the number of reported nodules after seeing the AI recommendations.

Correctness of Follow-Up Advice

Furthermore, to evaluate whether AI-CAD assistance resulted in improved clinical assessment, we analyzed whether the radiologists selected the correct follow-up advice more often with or without the AI recommendations. For each case, the correct follow-up according to the Fleischner criteria was retrospectively determined by 2 expert radiologists in consensus and used as reference follow-up advice. The follow-up recommendations provided by the radiologists were compared with the reference follow-up advice, and we analyzed whether AI assistance resulted in more accurate follow-up advice.

Statistical Analyses

Mental Model and Psychological Trust

Differences between the mental model scores and psychological trust scores of the radiologists at different time points were analyzed using the Wilcoxon signed rank test. Differences between the mental model scores and psychological trust scores of the groups with informative and reflective onboarding tutorials and of the groups with black box and explainable AI output were statistically analyzed using Mann-Whitney *U* tests. To control for heterogeneity, we tested whether radiologists' characteristics influenced the mental model scores and psychological trust scores at different time points and over time by performing multiple linear regression analyses.

Use of AI Recommendations

Multilevel logistic regression analyses were performed to assess whether the type of onboarding tutorial or level of explainability

of the AI output influenced the use of the AI recommendations and the radiologists' confidence in their assessments. To control for potential impact on the outcomes by other factors (exclusively the same nodules found by radiologists and AI model, change in number of reported nodules, age, gender, years of experience, how frequently they assess lungs on CT, how eager they are to try new information technologies, and how frequently they used computer-aided detection tools), these factors were included in the multilevel regression analyses as well. The same analysis scheme was used for all multilevel logistic regression analyses. First, an empty model was run to identify the variance at the individual level. The second regression analysis also considered the variants of onboarding tutorials and AI output. Third, whether the same nodules were found by AI and the radiologist exclusively and whether they made changes in the number of reported nodules were added. The final analysis also included different CT scans and radiologists' characteristics.

A P value of $<.05$ was considered statistically significant. All analyses were performed using Stata (version 17; StataCorp).

Ethical Considerations

This study was approved by the Internal Committee for Biomedical Experiments of Philips (number ICBE-S-000204) and conducted in accordance with the Declaration of Helsinki (as revised in 2013). Written informed consent was obtained from the participating clinicians.

Results

Participants

In total, 20 physicians from 7 Dutch hospitals participated in this study. Of the 20 participants, 16 were radiologists (median 10.5, range 1-32 years of experience as a specialist), 1 was a nuclear radiologist (2 years of experience in assessing lung CT scans), and 3 were radiology residents (median 2, range 1-5 years of residency). Of the 16 radiologists, 8 (50%) specialized in thoracic radiology. The male-to-female ratio was 50:50. Of the participants, 25% (5/20) were aged between 26 and 35 years, 35% (7/20) were aged between 36 and 45 years, 20% (4/20)

were aged between 46 and 55 years, and 20% (4/20) were aged between 56 and 65 years.

Mental Model and Psychological Trust

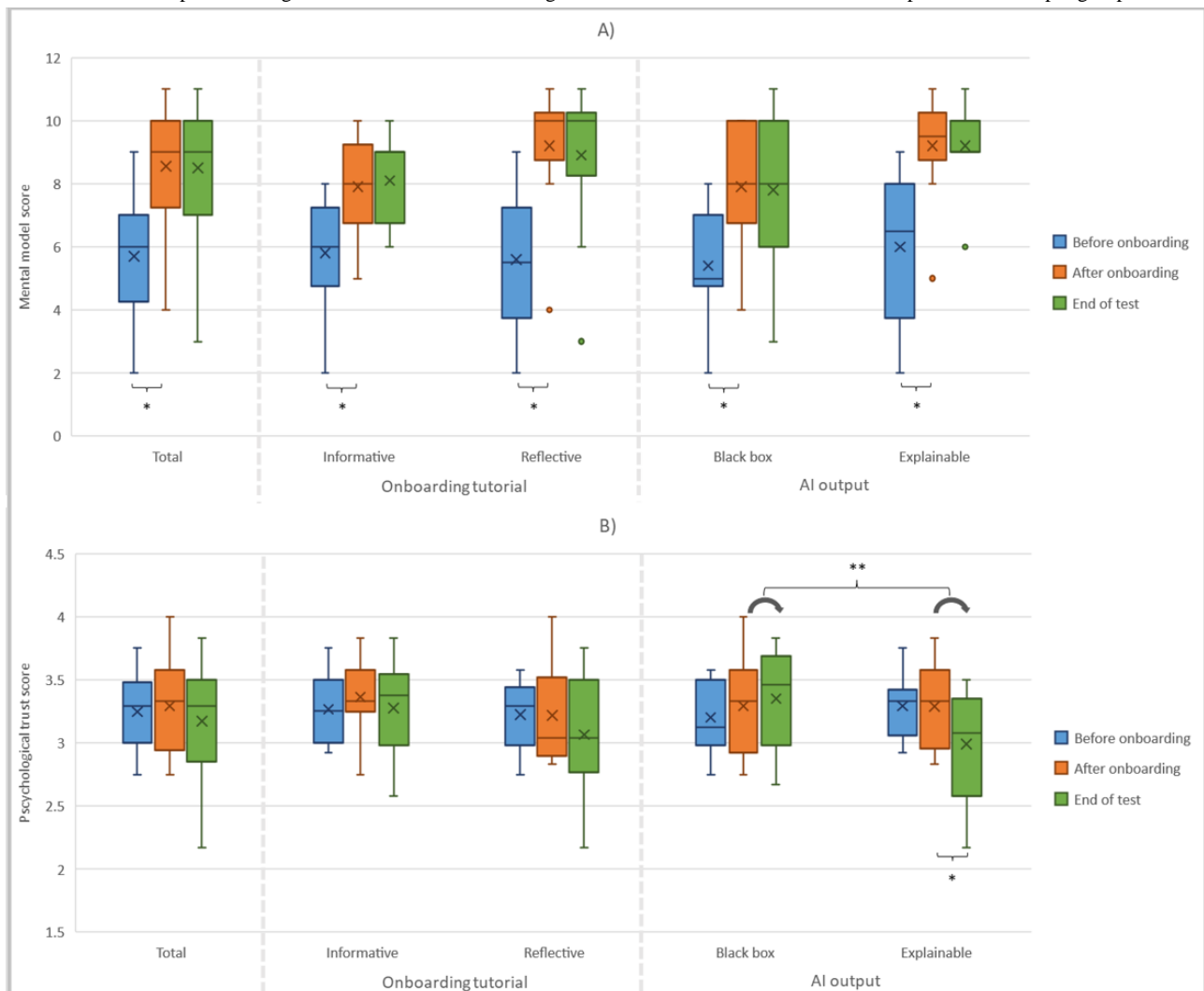
Figure 5 presents the mental model and psychological trust scores before onboarding, after onboarding, and at the end of the test. These scores were shown for all radiologists together and for the 2 variations of the onboarding tutorials and AI output separately.

After onboarding, the mental model score of the radiologists was significantly higher than that before onboarding ($P<.001$). The mean scores were 5.7 (SD 2.0) before onboarding and 8.6 (SD 1.9) after onboarding, which supports hypothesis 3a. Both informative ($P=.01$) and reflective ($P=.01$) onboarding resulted in significantly higher mental model scores. These improvements did not significantly differ between the groups; therefore, hypothesis 4a is not supported. At the end of the test, the mental model scores did not differ significantly from the scores after onboarding in any of the groups, which does not support hypothesis 1a and hypothesis 6a.

Considering all radiologists together, the psychological trust scores did not change significantly over time; therefore, hypotheses 1b and 3b are not supported. Between the 2 variations of onboarding tutorials, no significant differences in psychological trust scores were observed, and therefore, hypothesis 4b is not supported. In the group that received explainable AI output, psychological trust at the end of the test was significantly lower than that after onboarding ($P=.02$), which interestingly contradicts hypothesis 6b. In the group that received black box AI output, there was no significant change in psychological trust. Changes in psychological trust scores between after onboarding and at the end of the test were significantly different between the black box output and explainable AI output groups ($P=.03$). All P values can be found in [Multimedia Appendix 5](#).

None of the tested characteristics of radiologists significantly predicted the mental model scores or the psychological trust scores at the different time points nor did they significantly predict the changes over time.

Figure 5. Boxplot showing the (A) mental model scores and (B) psychological trust scores before and after onboarding and at the end of the test using either informative or reflective onboarding tutorials and either black box or explainable artificial intelligence (AI) output. The cross shows the mean value; the horizontal line inside the box indicates the median value; the lower and higher boundaries of the box indicate the first and third quartiles; the whiskers indicate the minimum and maximum values; and outliers are indicated by colored dots. Only significant differences are mentioned. *Significant difference between time points. **Significant difference in the change over time between the black box and explainable AI output groups.



Use of AI Recommendations

After viewing the AI outcomes, the radiologists adjusted their found nodules in 27 of 140 assessments, their estimated probability of malignancy in 32 of 140 assessments, and their follow-up advice in 12 of 140 assessments (Figure 6). Radiologists predominantly added nodules (23 of 27 changed cases), increased the probability of malignancy (24 of 32 changed cases), and shortened the recommended follow-up period (eg, from CT at 6-12 months to CT at 3-6 months; 8 of 12 changed cases). The empty model, which included no predictor variables, revealed that regarding whether radiologists made changes, approximately 3% of the variance in the outcome variable was attributable to differences between radiologists. For changes in malignancy prediction and follow-up advice, this attributable variance was approximately 20% and 7%, respectively. This indicates that there is some variability in the outcome, which can be explained by the individual radiologists. Radiologists' assessments were not significantly impacted by the type of onboarding tutorial or by the type of AI output; therefore, hypotheses 5 and 7 are not supported. All outcomes

of the multilevel regression analyses can be found in [Multimedia Appendix 6](#).

At all levels of assessment, radiologists' confidence in the assessments (n=140) predominantly increased after viewing the AI-CAD recommendations (in found nodules [75/82, 91%] of all changed assessments, in malignancy probability [42/50, 84%], in follow-up advice [22/28, 79%]; Figure 7), which supports hypothesis 2. The multilevel regression analysis revealed that in the empty model without predictor variables, approximately 20% of the total variance in the changed confidence in detected nodules was attributed to differences between radiologists. Regarding the changed confidence in malignancy prediction and follow-up advice, this attribution of the total variance was 10% and 7%, respectively. The radiologists' confidence in their assessments was not significantly affected by the type of onboarding tutorial but was affected by the type of AI output after controlling for whether the AI model found the same or different nodules as the radiologist without AI assistance (first model: $\beta=0.143$; $P=.16$; second model: $\beta=0.167$; $P=.04$; third model: $\beta=0.207$; $P=.02$).

See [Multimedia Appendix 6](#) for all outcomes of the multilevel regression analyses.

Figure 6. Bar graph showing the changes in the radiologist’s computed tomography assessments; (A) Reported nodules, (B) Malignancy probability, (C) Follow-up advice after viewing the recommendations from the artificial intelligence–based computer-aided detection or diagnosis using either informative or reflective onboarding tutorials, and either black box or explainable artificial intelligence (AI) output. No significant differences between the onboarding and AI output groups resulted from the multilevel regression analyses.

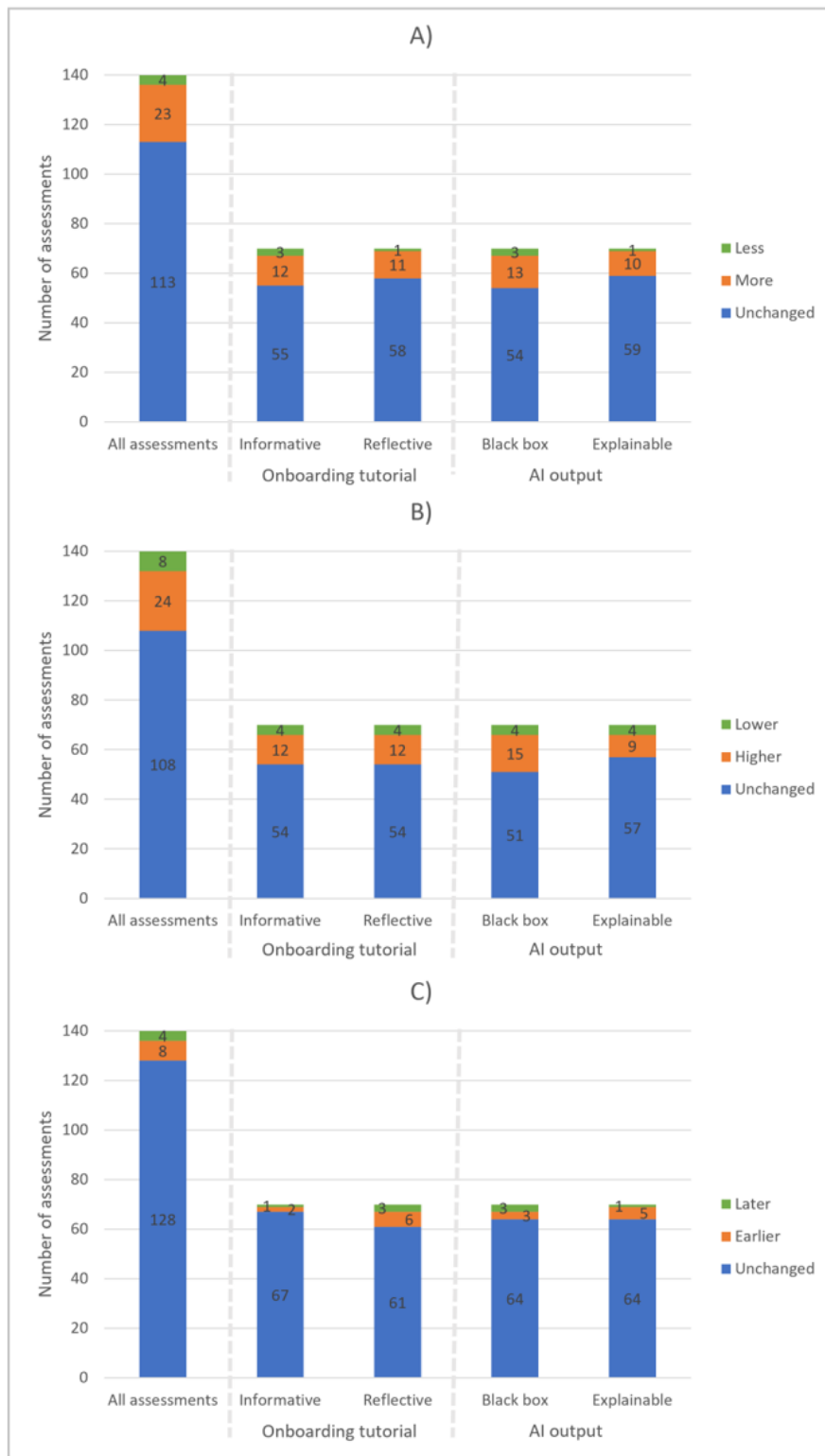
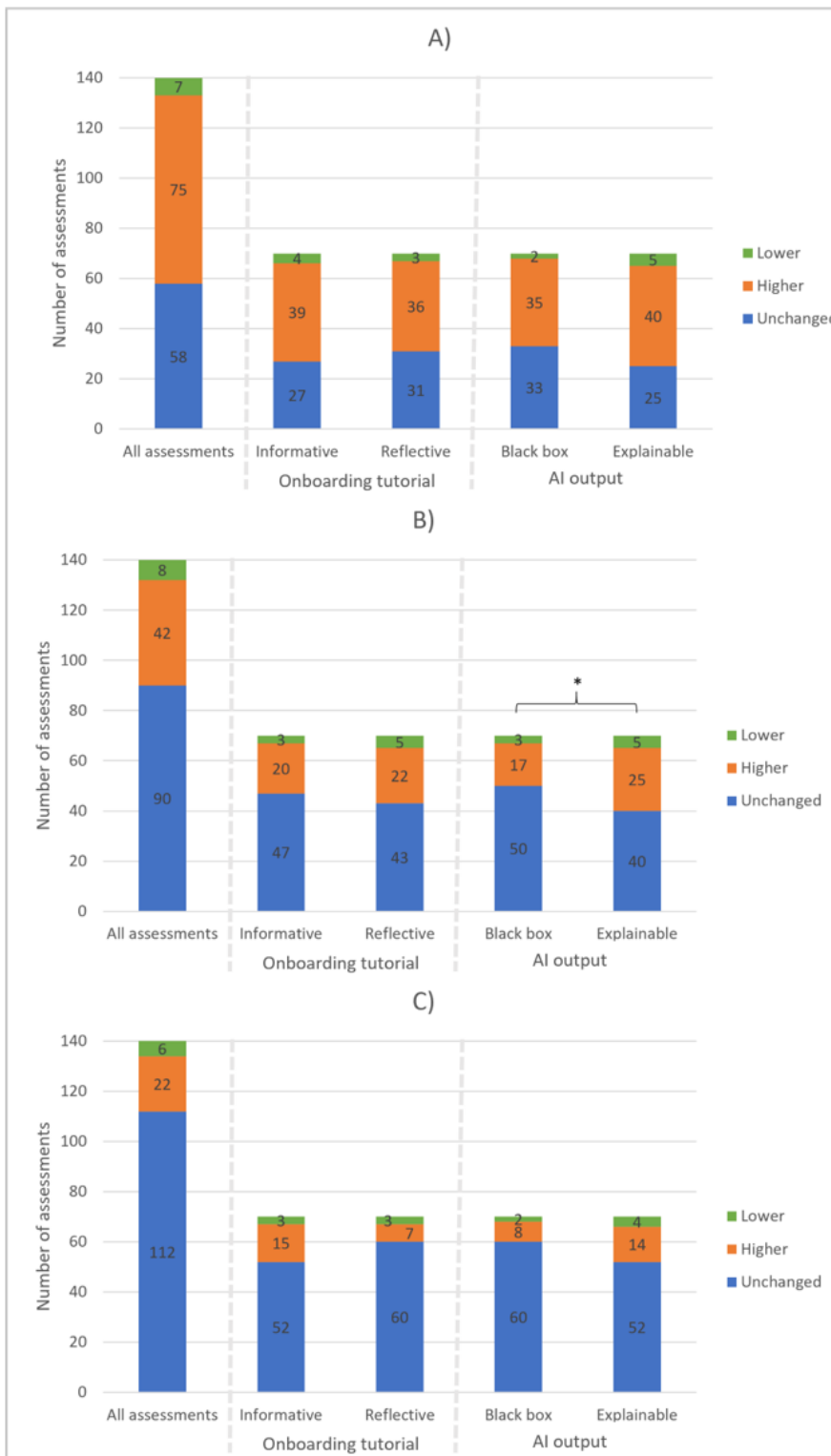


Figure 7. Bar graph showing the changes in the radiologist’s confidence in their assessments; (A) Confidence reported nodules, (B) Confidence malignancy probability, (C) Confidence follow-up advice after viewing the recommendations from the artificial intelligence–based computer-aided detection or diagnosis using either informative or reflective onboarding tutorials, and either black box or explainable artificial intelligence (AI) output. *The multilevel regression analysis showed a significant difference between the 2 groups according to the number of changed radiologists’ confidence (orange+green) in their assessment after using the artificial intelligence–based computer-aided detection or diagnosis system.



Secondary Outcomes

Post Hoc Analyses Regarding the Use of AI Recommendations

In 26 of 140 assessments, the same nodules exclusively had been found by the AI model and the unassisted radiologist. In these cases, radiologists changed the number of nodules less frequently than when different nodules had been found (second model: $\beta=-0.245$; $P=.003$ third model: $\beta=-0.437$; $P=.001$; [Multimedia Appendix 6](#)).

In 27 of 140 assessments, radiologists changed the number of nodules when using AI assistance. In the cases in which the radiologists did not change the number of nodules, the radiologists' confidence in their malignancy prediction changed more often, mostly increased, than in the cases in which the radiologists did change the number of found nodules (second model: $\beta=0.369$; $P<.001$; third model: $\beta=0.283$; $P=.001$; [Multimedia Appendix 6](#)). Whether the number of nodules was changed also significantly influenced radiologists' confidence in their follow-up advice, but this was probably related to some

radiologists' characteristics, as this effect disappeared after controlling for such characteristics (second model: $\beta=0.277$; $P=.02$; third model: $\beta=0.154$; $P=.23$).

Correctness Follow-Up Advice

Without AI assistance, the radiologists provided the correct follow-up advice according to the Fleischner criteria in 94 of 140 assessments ([Table 1](#)). Mostly, the correct follow-up advice was provided for CT cases 1, 3, 5, and 7, whereas most of the incorrect follow-up advice concerned CT cases 2, 4, and 6. With AI assistance, radiologists provided correct follow-up advice in 100 of 140 assessments. In 12 cases, the follow-up advice was changed after viewing the AI results. In 7 of these 12 cases, correct follow-up was provided after seeing the AI results. In 1 case, correct follow-up advice that was given initially was changed to incorrect follow-up advice after seeing the AI results. In 3 cases, the changed follow-up advice was still not correct but closer to the correct follow-up advice, and in the remaining case, the changed follow-up advice was further from the correct follow-up advice.

Table 1. Correct follow-up advice provided by the radiologists.

CT ^a cases (number of assessments)	All (n=140)	CT1 (n=20)	CT2 (n=20)	CT3 (n=20)	CT4 (n=20)	CT5 (n=20)	CT6 (n=20)	CT7 (n=20)
Correct follow-up advice given without AI ^b assistance, n (%)	94 (67)	20 (100)	6 (30)	17 (85)	6 (30)	20 (100)	7 (35)	18 (90)
Correct follow-up advice given with AI assistance, n (%)	100 (71)	20 (100)	7 (35)	17 (85)	7 (35)	20 (100)	9 (45)	20 (100)
Changed follow-up advice after using AI assistance, n (%)	12 (9)	0 (0)	2 (10)	1 (5)	5 (25)	0 (0)	2 (10)	2 (10)
Wrong→correct	7 (58)	0 (0)	1 (50)	0 (0)	2 (40)	0 (0)	2 (100)	2 (100)
Wrong→better (still wrong, but closer to correct follow-up)	3 (25)	0 (0)	1 (50)	1 (100)	1 (20)	0 (0)	0 (0)	0 (0)
Wrong→worse (still wrong, even further from correct follow-up)	1 (8)	0 (0)	0 (0)	0 (0)	1 (20)	0 (0)	0 (0)	0 (0)
Correct→wrong	1 (8)	0 (0)	0 (0)	0 (0)	1 (20)	0 (0)	0 (0)	0 (0)

^aCT: computed tomography.

^bAI: artificial intelligence.

Discussion

Principal Findings

Our study demonstrated that onboarding is of great importance because the radiologists' mental model of the AI-CAD system was significantly more accurate after onboarding. This finding implies that after onboarding, radiologists had a better understanding of the capabilities and limitations of the AI-CAD system, which is important for using the AI recommendations correctly. In addition, the importance of onboarding was emphasized by the fact that the mental model did not become more accurate through the actual use of the AI-CAD system. A study by Lam Shin Cheung et al [45] supports the need for onboarding.

We hypothesized that reflective onboarding would result in a more appropriate level of trust than informative onboarding, as radiologists in the reflective onboarding group were triggered to actively engage in cognitive reflection and receive feedback on their mental model. However, this hypothesis was not supported because the increases in mental model scores of

radiologists in the reflective onboarding group did not significantly differ from those in the informative onboarding group. This unexpected finding might be explained by the high level of clarity of the explanations provided during both informative and reflective onboarding, because of which the reflection had no significant added value. Alternatively, participating radiologists might possess a natural tendency to engage in cognitive reflection even if the system does not actively trigger them to do so.

Another unexpected finding was that explainable AI output resulted in a significant decrease in psychological trust ($P=.02$) during the use of the AI-CAD system for assessing the 7 CT scans, which was not the case in the group that received black box AI output ([Figure 5](#)). Apparently, users can become insecure about the reliability of AI-CAD when they receive explanations. On the basis of feedback from the participating radiologists, we know that some radiologists observed that the AI-CAD system provided different malignancy predictions for similar nodules with the same visual characteristics provided such as size and morphology. These discrepancies raised questions about why

nodules with similar characteristics had different malignancy probabilities. In fact, this key aspect still felt like a black box to the participants. Apparently, providing more transparency, which enables radiologists to observe inconsistencies in the AI predictions, can decrease the radiologists' trust in the AI-CAD system. However, this decrease in trust might be appropriate because the AI model's performance might be suboptimal and inconsistent.

In many CT assessments, the radiologists did not make any changes in their assessments after seeing the AI recommendations. However, this does not necessarily mean that the radiologist did not trust the AI-CAD system. There can be several reasons for making no changes. First, the AI recommendations can be exactly the same as the radiologists' assessments. Second, radiologists may disagree with the AI recommendations, which may be appropriate because the AI model also makes mistakes. Third, concerning malignancy prediction and follow-up advice, the AI recommendations may not impact the assessments, whereas the radiologists do agree with the AI recommendations. For instance, the AI model might find an extra nodule; however, if another larger and more suspicious nodule was already detected, the extra nodule does not impact the radiologist's malignancy risk prediction at the patient level or the follow-up recommendation.

Another important finding is that radiologists became more confident in their assessments after using the AI recommendations. This change might be explained by the fact that the AI-CAD system provides an extra check, which reduces the likelihood of nodules being overlooked. Hence, it provides radiologists with a sense of safety that increases their confidence, regardless of whether they agree with the AI output.

The follow-up advice was adjusted by the radiologists after viewing the AI results in only 12 of 140 assessments, whereas the number of observed nodules and the malignancy probabilities were changed more often (27/140, 19.3% assessments and 32/140, 22.9% assessments, respectively). This finding can be explained by the fact that follow-up advice is predominantly affected by the most suspicious nodule. Consequently, an AI-CAD finding of an additional small nodule while a large suspicious nodule had already been detected by the radiologist did not impact the radiologist's follow-up advice. Of the 3 assessment levels, follow-up advice is clinically most relevant. When the follow-up advice was adjusted, it was mostly changed to a shorter follow-up period (8/12, 67% assessments; eg, from CT at 6-12 months to CT at 3-6 months). This finding indicates that, owing to the AI recommendations, radiologists tended to be more careful and took fewer risks in their follow-up advice. For this study, earlier follow-up was appropriate as all CT scans showed cancer cases, but in clinical practice, it can be questionable whether being more careful and taking fewer risks in the follow-up advice is always desirable because it may increase the health care costs. Therefore, it is of great importance to study the cost-effectiveness of AI-CAD systems.

Secondary Findings

Confidence in malignancy prediction was significantly more frequently changed when the radiologist did not change their number of nodules after viewing the AI recommendations

([Multimedia Appendix 6](#)). This might be caused by the malignancy prediction provided by the AI-CAD system of nodules that they also found themselves. The radiologist might become more convinced whether a case is malignant or benign based on this AI-CAD malignancy recommendation.

This study also demonstrates the importance of applying a user-centered design process to achieve appropriate use of the AI-CAD system. This is lacking in many studies and applications [46]. Radiologists indicated in their feedback that the PPV and NPV were difficult to interpret. Therefore, different visualizations of model confidence might be more appropriate, such as using only bar graphs. Furthermore, radiologists mentioned that some extra functionalities that radiologists use in clinical practice for lung assessment need to be implemented in the prototype, such as multiplanar reconstruction and maximum intensity projection, underlining the need for tight integration of AI into the radiologist routine workstations. In addition, they mentioned that during onboarding, they would like to receive more information on AI model training and validation, including the data sets used and ground truth definition, which should therefore be added to the onboarding prototype. This need is in line with the findings of Cai et al [31], who explored the information needs for onboarding for AI-CAD in pathology. Ashoori and Weisz [43] mentioned that information on AI model training and testing is important for radiologists' trust in AI-CAD systems. Radiologists' feedback needs to be incorporated to achieve the AI-CAD system that fully meets radiologists' needs.

Limitations and Future Perspectives

This study had several limitations. First, this study was not fully representative of the clinical situation. Owing to time constraints, we specifically asked the radiologists not to assess the entire case but to focus on the component task of lung nodule assessment. Therefore, radiologists were aware that lung nodule assessment was important, which is representative for CT scans acquired because of pulmonary complaints but not for scans with incidental lung nodules. In addition, this study exclusively included scans of cancer cases, which differs from clinical practice, in which scans may also show no nodules and solely benign nodules. However, the data set with cancer cases was appropriate for our research goals.

Second, in the current prototype, the explainable AI output was simulated post hoc. There is an increasingly louder call to build causal models in the medical domain where the cost of failure is high, allowing the clinician to verify the causal chain of effects of clinically validated features on the model prediction. However, such inherently interpretable models are currently the exception rather than mainstream practice [47]. In this study, we focused on the current state of medical practice, where, if at all, most post hoc explainability techniques are used to improve interpretability. Importantly, post hoc techniques come at the expense of the validity of the relationship between post hoc explanations and model prediction. In fact, what appears to an end user as an explanation might not convey why the black box predicted what it did [48]. In this study, we were interested in the effect of a widespread approach to explain user trust and decision-making in a medical context. In addition, although

simulating explainable AI output is very useful in the early stages of AI-CAD system development [33,34], having fully functioning AI models would further add to the realism of the test. Furthermore, it would be valuable if the algorithm can provide the extent to which each nodule characteristic contributed to malignancy prediction. In addition, PPV and NPV computed at the patient level were applied at the nodule level.

Third, this study included only 20 radiologists and 7 CT scans, which need to be scaled up to have sufficient power to be able to detect smaller effect sizes. In this pilot study, this limitation was accepted to make the test less time-consuming for the participating radiologists and to postpone larger samples after at least some evidence of larger effects in this context could be established. During case selection for this study, we aimed to collect a mix of relatively easy and more challenging cases, which worked well, considering the number of correct follow-up recommendations in Table 1. In a future large-scale study, it would be advisable to use a clinically representative data set to prevent the impact of selection bias. Testing on a larger scale is also required to analyze what radiologists do with FP findings

and how these findings affect their trust in the AI-CAD. It is interesting to assess which types of FP findings are recognized by radiologists. Furthermore, it is useful to analyze whether changes in the number of observed nodules and in malignancy probability are correct based on a reference standard defined by expert radiologists and pathology. This is important because of automation bias, implying that radiologists rely too much on the AI recommendations, has to be prevented [40,49].

Conclusions

When clinical decision support systems are implemented, clinicians should receive careful onboarding that gives them a better understanding of the capabilities and limitations of the AI-CAD system. This understanding contributes to appropriate trust in the AI system, which is important when AI systems are used in clinical practice. Providing more AI output transparency, which enables clinicians to observe inconsistencies in the AI recommendations, can decrease clinicians' trust in the AI-CAD system. AI recommendations frequently increased radiologists' confidence in their assessments, even if they did not fully agree with these recommendations.

Acknowledgments

The members of the e/MTIC Oncology group are Fons van der Sommen, Joost Nederend, Misha D P Luyer, Mathias Funk, Jon R Pluyter, Igor Jacobs, , Dimitrios Mavroeidis, Chris C P Snijders, Susan Hommerson, Lotte J S Ewals, Mark Ramaekers, Kasper van der Wulp, Christiaan G A Viviers, Terese A E Hellström, Nick H C Ruijs, Ning Fang and Victoria Bruno.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Characteristics of computed tomography cases.

[[DOCX File , 17 KB - ai_v3i1e52211_app1.docx](#)]

Multimedia Appendix 2

Positive predictive value and negative predictive value definitions.

[[DOCX File , 15 KB - ai_v3i1e52211_app2.docx](#)]

Multimedia Appendix 3

Experimental conditions.

[[DOCX File , 1239 KB - ai_v3i1e52211_app3.docx](#)]

Multimedia Appendix 4

Forms for measuring trust.

[[DOCX File , 24 KB - ai_v3i1e52211_app4.docx](#)]

Multimedia Appendix 5

Mental model and psychological trust.

[[DOCX File , 16 KB - ai_v3i1e52211_app5.docx](#)]

Multimedia Appendix 6

Use of artificial intelligence recommendations.

[[DOCX File , 22 KB - ai_v3i1e52211_app6.docx](#)]

References

<https://ai.jmir.org/2024/1/e52211>

JMIR AI 2024 | vol. 3 | e52211 | p.552
(page number not for citation purposes)

1. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global Cancer Statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin* 2021 May;71(3):209-249 [FREE Full text] [doi: [10.3322/caac.21660](https://doi.org/10.3322/caac.21660)] [Medline: [33538338](https://pubmed.ncbi.nlm.nih.gov/33538338/)]
2. Rubin GD. Lung nodule and cancer detection in computed tomography screening. *J Thorac Imaging* 2015 Mar;30(2):130-138 [FREE Full text] [doi: [10.1097/RTI.0000000000000140](https://doi.org/10.1097/RTI.0000000000000140)] [Medline: [25658477](https://pubmed.ncbi.nlm.nih.gov/25658477/)]
3. Del Ciello A, Franchi P, Contegiacomo A, Cicchetti G, Bonomo L, Larici AR. Missed lung cancer: when, where, and why? *Diagn Interv Radiol* 2017 Mar 01;23(2):118-126 [FREE Full text] [doi: [10.5152/dir.2016.16187](https://doi.org/10.5152/dir.2016.16187)] [Medline: [28206951](https://pubmed.ncbi.nlm.nih.gov/28206951/)]
4. Hossain R, Wu CC, de Groot PM, Carter BW, Gilman MD, Abbott GF. Missed lung cancer. *Radiol Clin North Am* 2018 May;56(3):365-375. [doi: [10.1016/j.rcl.2018.01.004](https://doi.org/10.1016/j.rcl.2018.01.004)] [Medline: [29622072](https://pubmed.ncbi.nlm.nih.gov/29622072/)]
5. Li F, Sone S, Abe H, MacMahon H, Armato SG, Doi K. Lung cancers missed at low-dose helical CT screening in a general population: comparison of clinical, histopathologic, and imaging findings. *Radiology* 2002 Dec;225(3):673-683. [doi: [10.1148/radiol.2253011375](https://doi.org/10.1148/radiol.2253011375)] [Medline: [12461245](https://pubmed.ncbi.nlm.nih.gov/12461245/)]
6. Horeweg N, Scholten ET, de Jong PA, van der Aalst CM, Weenink C, Lammers JJ, et al. Detection of lung cancer through low-dose CT screening (NELSON): a prespecified analysis of screening test performance and interval cancers. *Lancet Oncol* 2014 Nov;15(12):1342-1350. [doi: [10.1016/S1470-2045\(14\)70387-0](https://doi.org/10.1016/S1470-2045(14)70387-0)] [Medline: [25282284](https://pubmed.ncbi.nlm.nih.gov/25282284/)]
7. Firmino M, Angelo G, Morais H, Dantas MR, Valentim R. Computer-aided detection (CADe) and diagnosis (CADx) system for lung cancer with likelihood of malignancy. *Biomed Eng Online* 2016 Jan 06;15(1):2 [FREE Full text] [doi: [10.1186/s12938-015-0120-7](https://doi.org/10.1186/s12938-015-0120-7)] [Medline: [26759159](https://pubmed.ncbi.nlm.nih.gov/26759159/)]
8. Gu Y, Chi J, Liu J, Yang L, Zhang B, Yu D, et al. A survey of computer-aided diagnosis of lung nodules from CT scans using deep learning. *Comput Biol Med* 2021 Oct;137:104806. [doi: [10.1016/j.compbiomed.2021.104806](https://doi.org/10.1016/j.compbiomed.2021.104806)] [Medline: [34461501](https://pubmed.ncbi.nlm.nih.gov/34461501/)]
9. Xie Z. Towards single-phase single-stage detection of pulmonary nodules in chest CT imaging. arXiv Preprint posted online July 16, 2018. [FREE Full text] [doi: [10.48550/arXiv.1807.05972](https://doi.org/10.48550/arXiv.1807.05972)]
10. Wu Z, Wang F, Cao W, Qin C, Dong X, Yang Z, et al. Lung cancer risk prediction models based on pulmonary nodules: a systematic review. *Thoracic Cancer* 2022 Mar 08;13(5):664-677 [FREE Full text] [doi: [10.1111/1759-7714.14333](https://doi.org/10.1111/1759-7714.14333)] [Medline: [35137543](https://pubmed.ncbi.nlm.nih.gov/35137543/)]
11. Gu D, Liu G, Xue Z. On the performance of lung nodule detection, segmentation and classification. *Comput Med Imaging Graph* 2021 Apr;89:101886. [doi: [10.1016/j.compmedimag.2021.101886](https://doi.org/10.1016/j.compmedimag.2021.101886)] [Medline: [33706112](https://pubmed.ncbi.nlm.nih.gov/33706112/)]
12. Ewals LJS, van der Wulp K, van den Borne BEEM, Pluyter JR, Jacobs I, Mavroidis D, et al. The effects of artificial intelligence assistance on the radiologists' assessment of lung nodules on CT scans: a systematic review. *J Clin Med* 2023 May 18;12(10):3536 [FREE Full text] [doi: [10.3390/jcm12103536](https://doi.org/10.3390/jcm12103536)] [Medline: [37240643](https://pubmed.ncbi.nlm.nih.gov/37240643/)]
13. Jeyakumar T, Younus S, Zhang M, Clare M, Charow R, Karsan I, et al. Preparing for an artificial intelligence-enabled future: patient perspectives on engagement and health care professional training for adopting artificial intelligence technologies in health care settings. *JMIR Preprints*: e40973 Preprint posted online March 2, 2023. [FREE Full text] [doi: [10.2196/40973](https://doi.org/10.2196/40973)]
14. Castelo N, Bos MW, Lehmann DR. Task-dependent algorithm aversion. *J Mark Res* 2019 Jul 15;56(5):809-825. [doi: [10.1177/0022243719851788](https://doi.org/10.1177/0022243719851788)]
15. Glikson E, Woolley AW. Human trust in artificial intelligence: review of empirical research. *Acad Manag Ann* 2020 Jul;14(2):627-660. [doi: [10.5465/annals.2018.0057](https://doi.org/10.5465/annals.2018.0057)]
16. Lee JD, See KA. Trust in automation: designing for appropriate reliance. *Hum Factors* 2004;46(1):50-80. [doi: [10.1518/hfes.46.1.50_30392](https://doi.org/10.1518/hfes.46.1.50_30392)] [Medline: [15151155](https://pubmed.ncbi.nlm.nih.gov/15151155/)]
17. Siau K, Wang W. Building trust in artificial intelligence, machine learning, and robotics. *Cut Bus Technol J* 2018;31(2):47-53 [FREE Full text]
18. Ezer N, Bruni S, Cai Y, Hepenstal SJ, Miller CA, Schmorow DD. Trust engineering for human-AI teams. *Proc Hum Factors Ergon Soc Annu Meet* 2019 Nov 20;63(1):322-326. [doi: [10.1177/1071181319631264](https://doi.org/10.1177/1071181319631264)]
19. Martínez-Torres MR, Díaz-Fernández MC, Toral SL, Barrero F. The moderating role of prior experience in technological acceptance models for ubiquitous computing services in urban environments. *Technol Forecast Soc Change* 2015 Feb;91:146-160. [doi: [10.1016/j.techfore.2014.02.004](https://doi.org/10.1016/j.techfore.2014.02.004)]
20. Schaefer KE, Chen JYC, Szalma JL, Hancock PA. A meta-analysis of factors influencing the development of trust in automation: implications for understanding autonomy in future systems. *Hum Factors* 2016 May 22;58(3):377-400. [doi: [10.1177/0018720816634228](https://doi.org/10.1177/0018720816634228)] [Medline: [27005902](https://pubmed.ncbi.nlm.nih.gov/27005902/)]
21. Jorritsma W, Cnossen F, van Ooijen PMA. Improving the radiologist-CAD interaction: designing for appropriate trust. *Clin Radiol* 2015 Feb;70(2):115-122. [doi: [10.1016/j.crad.2014.09.017](https://doi.org/10.1016/j.crad.2014.09.017)] [Medline: [25459198](https://pubmed.ncbi.nlm.nih.gov/25459198/)]
22. Muir BM. Trust between humans and machines, and the design of decision aids. *Int J Man Mach Stud* 1987 Nov;27(5-6):327-339. [doi: [10.1016/S0020-7373\(87\)80013-5](https://doi.org/10.1016/S0020-7373(87)80013-5)]
23. Meyer J, Lee JD. Trust, reliance, and compliance. In: Lee JD, Kirlik A, editors. *The Oxford Handbook of Cognitive Engineering*. Oxford, UK: Oxford Academic; 2013:109-124.
24. Endsley MR. Situation models: an avenue to the modeling of mental models. *Proc Hum Factors Ergon Soc Annu Meet* 2000 Jul 01;44(1):61-64. [doi: [10.1177/154193120004400117](https://doi.org/10.1177/154193120004400117)]

25. Collins MG, Juvina I. Trust miscalibration is sometimes necessary: an empirical study and a computational model. *Front Psychol* 2021 Aug 10;12:690089 [FREE Full text] [doi: [10.3389/fpsyg.2021.690089](https://doi.org/10.3389/fpsyg.2021.690089)] [Medline: [34447334](https://pubmed.ncbi.nlm.nih.gov/34447334/)]
26. Deutsch M. The effect of motivational orientation upon trust and suspicion. *Hum Relat* 1960 May;13(2):123-139. [doi: [10.1177/001872676001300202](https://doi.org/10.1177/001872676001300202)]
27. Bansal G, Nushi B, Kamar E, Weld DS, Lasecki WS, Horvitz E. Updates in human-AI teams: understanding and addressing the performance/compatibility tradeoff. *Proc AAAI Conf Artif Intell* 2019 Jul 17;33(01):2429-2437. [doi: [10.1609/aaai.v33i01.33012429](https://doi.org/10.1609/aaai.v33i01.33012429)]
28. Madsen M, Gregor S. Measuring human-computer trust. Central Queensland University. 2000. URL: <https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=b8eda9593fbc63b7ced1866853d9622737533a2> [accessed 2024-02-17]
29. Zavolokina L, Zani N, Schwabe G. Designing for trust in blockchain platforms. *IEEE Trans Eng Manage* 2023 Mar;70(3):849-863. [doi: [10.1109/tem.2020.3015359](https://doi.org/10.1109/tem.2020.3015359)]
30. Lee JD, Moray N. Trust, self-confidence, and operators' adaptation to automation. *Int J Hum Comput Stud* 1994 Jan;40(1):153-184. [doi: [10.1006/ijhc.1994.1007](https://doi.org/10.1006/ijhc.1994.1007)]
31. Cai CJ, Winter S, Steiner D, Wilcox L, Terry M. "Hello AI": uncovering the onboarding needs of medical practitioners for human-ai collaborative decision-making. *Proc ACM Hum Comput Interact* 2019 Nov 07;3:1-24. [doi: [10.1145/3359206](https://doi.org/10.1145/3359206)]
32. Mann K, Gordon J, MacLeod A. Reflection and reflective practice in health professions education: a systematic review. *Adv Health Sci Educ Theory Pract* 2009 Oct 23;14(4):595-621. [doi: [10.1007/s10459-007-9090-2](https://doi.org/10.1007/s10459-007-9090-2)] [Medline: [18034364](https://pubmed.ncbi.nlm.nih.gov/18034364/)]
33. Li AC, Kannry JL, Kushniruk A, Chrimes D, McGinn TG, Edonyabo D, et al. Integrating usability testing and think-aloud protocol analysis with "near-live" clinical simulations in evaluating clinical decision support. *Int J Med Inform* 2012 Nov;81(11):761-772. [doi: [10.1016/j.ijmedinf.2012.02.009](https://doi.org/10.1016/j.ijmedinf.2012.02.009)] [Medline: [22456088](https://pubmed.ncbi.nlm.nih.gov/22456088/)]
34. Matthiesen S, Diederichsen SZ, Hansen MKH, Villumsen C, Lassen MCH, Jacobsen PK, et al. Clinician preimplementation perspectives of a decision-support tool for the prediction of cardiac arrhythmia based on machine learning: near-live feasibility and qualitative study. *JMIR Hum Factors* 2021 Nov 26;8(4):e26964 [FREE Full text] [doi: [10.2196/26964](https://doi.org/10.2196/26964)] [Medline: [34842528](https://pubmed.ncbi.nlm.nih.gov/34842528/)]
35. Trajanovski S, Mavroeidis D, Swisher CL, Gebre BG, Veeling BS, Wiemker R, et al. Towards radiologist-level cancer risk assessment in CT lung screening using deep learning. *Comput Med Imaging Graph* 2021 Jun;90:101883. [doi: [10.1016/j.compmedimag.2021.101883](https://doi.org/10.1016/j.compmedimag.2021.101883)] [Medline: [33895622](https://pubmed.ncbi.nlm.nih.gov/33895622/)]
36. Liao F, Liang M, Li Z, Hu X, Song S. Evaluate the malignancy of pulmonary nodules using the 3-D deep leaky noisy-OR network. *IEEE Trans Neural Netw Learn Syst* 2019 Nov;30(11):3484-3495. [doi: [10.1109/TNNLS.2019.2892409](https://doi.org/10.1109/TNNLS.2019.2892409)] [Medline: [30794190](https://pubmed.ncbi.nlm.nih.gov/30794190/)]
37. National Lung Screening Trial Research Team, Aberle DR, Adams AM, Berg CD, Black WC, Clapp JD, et al. Reduced lung-cancer mortality with low-dose computed tomographic screening. *N Engl J Med* 2011 Aug 04;365(5):395-409 [FREE Full text] [doi: [10.1056/NEJMoal102873](https://doi.org/10.1056/NEJMoal102873)] [Medline: [21714641](https://pubmed.ncbi.nlm.nih.gov/21714641/)]
38. Leibig C, Allken V, Ayhan MS, Berens P, Wahl S. Leveraging uncertainty information from deep neural networks for disease detection. *Sci Rep* 2017 Dec 19;7(1):17816 [FREE Full text] [doi: [10.1038/s41598-017-17876-z](https://doi.org/10.1038/s41598-017-17876-z)] [Medline: [29259224](https://pubmed.ncbi.nlm.nih.gov/29259224/)]
39. Guo C, Pleiss G, Sun Y, Weinberger KQ. On calibration of modern neural networks. arXiv :1-14 Preprint posted online June 14, 2017. [FREE Full text] [doi: [10.1007/978-3-642-16712-6_101180](https://doi.org/10.1007/978-3-642-16712-6_101180)]
40. Rezazade Mehrizi MH, Mol F, Peter M, Ranschaert E, Dos Santos DP, Shahidi R, et al. The impact of AI suggestions on radiologists' decisions: a pilot study of explainability and attitudinal priming interventions in mammography examination. *Sci Rep* 2023 Jun 07;13(1):9230 [FREE Full text] [doi: [10.1038/s41598-023-36435-3](https://doi.org/10.1038/s41598-023-36435-3)] [Medline: [37286665](https://pubmed.ncbi.nlm.nih.gov/37286665/)]
41. Butz AM, Kaltenhauser A, Eiband M. The expert of Oz: a two-sided study paradigm for intelligent systems. In: *Proceedings of the 2020 ACM Designing Interactive Systems Conference*. 2020 Presented at: DIS' 20; July 6-10, 2020; Eindhoven, The Netherlands p. 269-273 URL: <https://dl.acm.org/doi/10.1145/3393914.3395874> [doi: [10.1145/3393914.3395874](https://doi.org/10.1145/3393914.3395874)]
42. Rosner B. *Fundamentals of Biostatistics*. Pacific Grove, CA: Duxbury Press; 2011.
43. Ashoori M, Weisz JD. In AI We trust? Factors that influence trustworthiness of AI-infused decision-making processes. arXiv :1-10 Preprint posted online December 5, 2019. [FREE Full text] [doi: [10.48550/arXiv.1912.02675](https://doi.org/10.48550/arXiv.1912.02675)]
44. MacMahon H, Naidich DP, Goo JM, Lee KS, Leung AN, Mayo JR, et al. Guidelines for management of incidental pulmonary nodules detected on CT images: from the Fleischner society 2017. *Radiology* 2017 Jul;284(1):228-243. [doi: [10.1148/radiol.2017161659](https://doi.org/10.1148/radiol.2017161659)] [Medline: [28240562](https://pubmed.ncbi.nlm.nih.gov/28240562/)]
45. Lam Shin Cheung J, Ali A, Abdalla M, Fine B. U"AI" testing: user interface and usability testing of a chest X-ray AI tool in a simulated real-world workflow. *Can Assoc Radiol J* 2023 May 02;74(2):314-325 [FREE Full text] [doi: [10.1177/08465371221131200](https://doi.org/10.1177/08465371221131200)] [Medline: [36189838](https://pubmed.ncbi.nlm.nih.gov/36189838/)]
46. Filice RW, Ratwani RM. The case for user-centered artificial intelligence in radiology. *Radiol Artif Intell* 2020 May 01;2(3):e190095 [FREE Full text] [doi: [10.1148/ryai.2020190095](https://doi.org/10.1148/ryai.2020190095)] [Medline: [33937824](https://pubmed.ncbi.nlm.nih.gov/33937824/)]
47. van Hartkamp M, Consoli S, Verhaegh W, Petkovic M, van de Stolpe A. Artificial intelligence in clinical health care applications: viewpoint. *Interact J Med Res* 2019 Apr 05;8(2):e12100 [FREE Full text] [doi: [10.2196/12100](https://doi.org/10.2196/12100)] [Medline: [30950806](https://pubmed.ncbi.nlm.nih.gov/30950806/)]

48. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat Mach Intell* 2019 May 13;1(5):206-215 [FREE Full text] [doi: [10.1038/s42256-019-0048-x](https://doi.org/10.1038/s42256-019-0048-x)] [Medline: [35603010](https://pubmed.ncbi.nlm.nih.gov/35603010/)]
49. Dratsch T, Chen X, Rezazade Mehrizi M, Kloeckner R, Mähringer-Kunz A, Püsken M, et al. Automation bias in mammography: the impact of artificial intelligence BI-RADS suggestions on reader performance. *Radiol* 2023 May;307(4):e222176. [doi: [10.1148/radiol.222176](https://doi.org/10.1148/radiol.222176)] [Medline: [37129490](https://pubmed.ncbi.nlm.nih.gov/37129490/)]

Abbreviations

AI: artificial intelligence

AI-CAD: artificial intelligence–based computer-aided detection or diagnosis

CT: computed tomography

FP: false-positive

NPV: negative predictive value

PPV: positive predictive value

Edited by K El Emam, B Malin; submitted 27.08.23; peer-reviewed by M Mehrizi, D Estevez Prado; comments to author 25.09.23; revised version received 14.11.23; accepted 03.02.24; published 13.03.24.

Please cite as:

Ewals LJS, Heesterbeek LJJ, Yu B, van der Wulp K, Mavroeidis D, Funk M, Snijders CCP, Jacobs I, Nederend J, Pluyter JR, e/MTIC Oncology group

The Impact of Expectation Management and Model Transparency on Radiologists' Trust and Utilization of AI Recommendations for Lung Nodule Assessment on Computed Tomography: Simulated Use Study

JMIR AI 2024;3:e52211

URL: <https://ai.jmir.org/2024/1/e52211>

doi: [10.2196/52211](https://doi.org/10.2196/52211)

PMID: [38875574](https://pubmed.ncbi.nlm.nih.gov/38875574/)

©Lotte J S Ewals, Lynn J J Heesterbeek, Bin Yu, Kasper van der Wulp, Dimitrios Mavroeidis, Mathias Funk, Chris C P Snijders, Igor Jacobs, Joost Nederend, Jon R Pluyter, e/MTIC Oncology group. Originally published in JMIR AI (<https://ai.jmir.org>), 13.03.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR AI, is properly cited. The complete bibliographic information, a link to the original publication on <https://www.ai.jmir.org/>, as well as this copyright and license information must be included.

Original Paper

Evaluation of Generative Language Models in Personalizing Medical Information: Instrument Validation Study

Aidin Spina¹, BSc; Saman Andalib¹, BSc; Daniel Flores^{1*}, BSc; Rishi Vermani^{1*}, BSc; Faris F Halaseh¹, BSc; Ariana M Nelson^{1,2}, MD

¹School of Medicine, University of California, Irvine, Irvine, CA, United States

²Department of Anesthesiology and Perioperative Care, University of California, Irvine, Irvine, CA, United States

*these authors contributed equally

Corresponding Author:

Aidin Spina, BSc

School of Medicine

University of California, Irvine

1001 Health Sciences Road

Irvine, CA, 92617

United States

Phone: 1 949 290 8347

Email: acspina@hs.uci.edu

Abstract

Background: Although uncertainties exist regarding implementation, artificial intelligence–driven generative language models (GLMs) have enormous potential in medicine. Deployment of GLMs could improve patient comprehension of clinical texts and improve low health literacy.

Objective: The goal of this study is to evaluate the potential of ChatGPT-3.5 and GPT-4 to tailor the complexity of medical information to patient-specific input education level, which is crucial if it is to serve as a tool in addressing low health literacy.

Methods: Input templates related to 2 prevalent chronic diseases—type II diabetes and hypertension—were designed. Each clinical vignette was adjusted for hypothetical patient education levels to evaluate output personalization. To assess the success of a GLM (GPT-3.5 and GPT-4) in tailoring output writing, the readability of pre- and posttransformation outputs were quantified using the Flesch reading ease score (FKRE) and the Flesch-Kincaid grade level (FKGL).

Results: Responses (n=80) were generated using GPT-3.5 and GPT-4 across 2 clinical vignettes. For GPT-3.5, FKRE means were 57.75 (SD 4.75), 51.28 (SD 5.14), 32.28 (SD 4.52), and 28.31 (SD 5.22) for 6th grade, 8th grade, high school, and bachelor's, respectively; FKGL mean scores were 9.08 (SD 0.90), 10.27 (SD 1.06), 13.4 (SD 0.80), and 13.74 (SD 1.18). GPT-3.5 only aligned with the prespecified education levels at the bachelor's degree. Conversely, GPT-4's FKRE mean scores were 74.54 (SD 2.6), 71.25 (SD 4.96), 47.61 (SD 6.13), and 13.71 (SD 5.77), with FKGL mean scores of 6.3 (SD 0.73), 6.7 (SD 1.11), 11.09 (SD 1.26), and 17.03 (SD 1.11) for the same respective education levels. GPT-4 met the target readability for all groups except the 6th-grade FKRE average. Both GLMs produced outputs with statistically significant differences ($P<.001$; 8th grade $P<.001$; high school $P<.001$; bachelors $P=.003$; FKGL: 6th grade $P=.001$; 8th grade $P<.001$; high school $P<.001$; bachelors $P<.001$) between mean FKRE and FKGL across input education levels.

Conclusions: GLMs can change the structure and readability of medical text outputs according to input-specified education. However, GLMs categorize input education designation into 3 broad tiers of output readability: easy (6th and 8th grade), medium (high school), and difficult (bachelor's degree). This is the first result to suggest that there are broader boundaries in the success of GLMs in output text simplification. Future research must establish how GLMs can reliably personalize medical texts to prespecified education levels to enable a broader impact on health care literacy.

(JMIR AI 2024;3:e54371) doi:[10.2196/54371](https://doi.org/10.2196/54371)

KEYWORDS

generative language model; GLM; artificial intelligence; AI; low health literacy; LHL; readability; GLMs; language model; language models; health literacy; understandable; understandability; knowledge translation; comprehension; generative; NLP;

natural language processing; reading level; reading levels; education; medical text; medical texts; medical information; health information

Introduction

Health literacy is critical for informed health care decisions. However, only 12% of Americans are considered to have proficient health literacy skills [1]. Low health literacy (LHL) is a limited ability to procure, process, and comprehend health information [2]. Importantly, patients with LHL have poorer health outcomes than those with higher health literacy [3]. Many interventions have been proposed and implemented to address health literacy disparities including community health fairs, increased number of primary care visits, and informational handouts [4]. Although the availability of community health fairs and on-demand primary care consultation is variable, the internet is widely accessible [5]. However, internet-derived health information has limitations. Specifically, accessing web-based information and navigating complex user interfaces results in information overload that can negate potential benefits for patients with LHL [6].

Artificial intelligence (AI)-driven chatbots use natural language processing to better interpret and respond to human-like prompts [7]. Generative language models (GLMs), such as ChatGPT (OpenAI), are now regularly used by consumers [8]. Despite the recent increase in the availability of GLMs, implementing AI as a patient education adjunct is not new [9]. Jayakumar et al [10] previously used AI to assist in patient medical education. The incorporation of AI-driven tools resulted in significantly improved decision quality and satisfaction among patients with knee osteoarthritis compared to patients who only received educational material [10]. Given the success of previous iterations of AI in patient education and decision-making, elucidating the potential role of a GLM in a similar capacity could be transformative as a resource to combat LHL [11]. These new tools for patient education can unlock methods for addressing health care concerns, such as pain perception, as illustrated by Sun et al [12], who used education to decrease perceived pain and facilitate recovery.

While there is ostensibly immense potential for this use of AI in health care, at this time, many questions remain, specifically about the accuracy and reproducibility of chatbot-generated medical content [7,13]. While content accuracy is a subject of further clinical discourse, this paper aims to explore the potential of GLMs in tailoring medical text to patient-specific characteristics such as education level. Previous research has evaluated the capability of ChatGPT to simplify a medical text and respond to hypothetical patient questions in various medical specialties [14-17]. In this study, we assessed the ability of ChatGPT versions 3.5 and 4 to transform text to suit a broad range of education levels, including 6th grade, 8th grade, 12th grade, and bachelor's degree. To elucidate this ability, we tested 2 common clinical scenarios: a patient learning about a diagnosis of diabetes mellitus (DM) or hypertension (HTN). The Flesch reading ease score (FKRE) and the Flesch-Kincaid grade level (FKGL) were implemented as outcome measures as both are clinically validated numeric text assessment tools. The FKRE and FKGL were originally developed to quantify readability

ease, with the FKGL being developed specifically for the US Navy in 1975 [18]. As indicated, scores are used commonly, for example, with the US Department of Defense using the FKRE to quantify the readability of its forms and documents [19]. These outcome measures have also frequently been implemented to assess the readability of clinical texts and have been frequently used for assessing outpatient resources [20,21]. Hence, this study explores GLMs as a potentially useful interface to combat LHL by personalizing medical information to a specific education level, as the complexity of clinician-provided and open-access medical information can often inhibit proper understanding. We hypothesized that ChatGPT would be able to successfully create outputs at different readabilities, with GPT-4 being more accurate than its predecessor model, GPT-3.5.

Methods

Overview

GPT-3.5 and GPT-4 were used for this study. Both models were developed using reinforcement learning from human feedback, which uses human-generated texts to prompt and train the GLM. This study used a standardized method to generate each input prompt, assess the readability of each output, and perform statistical analyses on the readability scores. This study focused exclusively on evaluating the capacity of GPT-3.5 and GPT-4 to generate outputs with targeted readability levels, without verifying the accuracy of the content produced.

Input Prompt Creation

A total of 2 input prompts were created that emulate common medical scenarios: DM and HTN. Pertinent information in each input prompt included patient demographics, chief concern at the time of presentation, a set of medical interventions to address the chief concern, and a sentence specifying the desired output (Figures S1-S4 in [Multimedia Appendix 1](#)).

Prespecified designation of input patient education level was the focus of this study. Previous research has demonstrated a significant correlation between educational attainment and health literacy, prompting us to use education level as a proxy for health literacy [22,23]. To explore the effect of changing the education level on the generated output, we repeatedly queried the same input prompt while only changing the designated education level of the patient. The education levels included a 6th-grade, 8th-grade, 12th-grade (high school graduate), and a university graduate (bachelor's degree). Starting at the 6th-grade level ensures alignment with standardized medical recommendations for reading levels of patient-facing materials, while the 8th grade represents the average reading level for an adult in the United States, aligning the study with broad public health guidelines [24-26]. High school graduates were evaluated to bridge the gap between middle and higher education, reflecting a common literacy standard, while the bachelor's degree level tests the GLM's ability to tailor complex health information for a more educated audience without unnecessary complexity. Specifically, the high school and bachelor's levels

of education were included as even highly educated people may have LHL due to the complexity of the medical text. Assessing whether GPT-3.5 and GPT-4 can customize outputs for these demographics is essential for validating their use as tools to potentially address LHL.

Generation of GPT-4 Outputs and Statistical Analysis

In total, 2 sets of input prompts (DM and HTN) were finalized and cataloged in spreadsheets. These input prompts were subsequently entered into GPT-3.5 and GPT-4. Each input scenario—such as GPT-3.5, DM, and 6th grade—was entered into a new conversation window, 5 separate times. All input scenarios were run 5 times to assess reproducibility and to attain statistical significance when comparing groups (Table 1). Next, each output was cataloged, placed into a standardized single-paragraph format, and entered into the readability calculator on Word (Microsoft Corp). The outputs were reformatted into a single paragraph to standardize readability scores, as variations in formatting can affect Word's ability to accurately measure readability. For this study, the FKRE and the FKGL values were calculated and subjected to statistical analysis [27]. Equations 1 and 2 show how each of these scores are calculated. The FKRE ranges from 0 to 100, with scores of 0 and 100 indicating texts of high and low reading complexity respectively (Table 2).

$$\text{FKRE} = 206.835 - 1.015 \times (\text{total words} \div \text{total sentences}) - 84.6 \times (\text{total syllables} \div \text{total words}) \quad (1)$$

$$\text{FKGL} = 0.39 \times (\text{total words} \div \text{total sentences}) + 11.8 \times (\text{total syllables} \div \text{total words}) - 15.59 \quad (2)$$

Single factor ANOVA was performed to determine if any significant differences existed between the education levels. Once ANOVA confirmed this statistically significant difference, each set of data was subjected to the Tukey multiple comparison post hoc analysis to evaluate differences between the means of each group within each scenario. Significance for all statistical analysis was set at $P < .05$. Single-factor ANOVA with the Tukey post hoc analysis was used because Shapiro-Wilk normality testing and Levene's test for equality of variances showed that the data did not violate the assumptions of normality or homogeneity of variances. Statistical analyses were performed for individual clinical scenarios and aggregated data. Unpaired 2-tailed t tests were also performed to determine the differences in functionality between GPT-3.5 and GPT-4 for both individual clinical scenarios and aggregated data across all 4 education levels. Finally, aggregated analysis was conducted to determine which education level led to outputs with the highest and lowest variation for FKRE and FKGL.

Table 1. Summary of scenarios organized by AI^a model, grade level, and clinical scenario (N=80).

Clinical scenarios and grade level	Number of scenarios, n (%)
GPT-4	
DM^b	
6th grade	5 (6)
8th grade	5 (6)
High school	5 (6)
Bachelor's	5 (6)
HTN^c	
6th grade	5 (6)
8th grade	5 (6)
High school	5 (6)
Bachelor's	5 (6)
GPT-3.5	
DM	
6th grade	5 (6)
8th grade	5 (6)
High school	5 (6)
Bachelor's	5 (6)
HTN	
6th grade	5 (6)
8th grade	5 (6)
High school	5 (6)
Bachelor's	5 (6)

^aAI: artificial intelligence.

^bDM: diabetes mellitus.

^cHTN: hypertension.

Table 2. Interpretation of the Flesch reading ease score based on the US grade level system.

Score	School level (US)	Description
10.0-0.0	Professional	Extremely difficult to read. Only suitable for university graduates
30.0-10.0	College graduate	Very difficult to read and comprehend
50.0-30.0	College	Difficult to read and comprehend
60.0-50.0	10th to 12th grade	Fairly difficult to read and comprehend
70.0-60.0	8th and 9th grade	"Plain English"
80.0-70.0	7th grade	Fairly easy to read and comprehend
90.0-80.0	6th grade	Easy to read and comprehend. Considered conversational English for speakers
100.0-90.0	5th grade	Extremely easy to read and comprehend

Ethical Considerations

No application was submitted for review board assessment because no human or animal participants participated directly or indirectly in this study. The University of California, Irvine Institutional Review Board does not require assessment of studies that do not directly or indirectly involve human or animal

participants. This study consisted solely of a quantitative evaluation of a GLM for text personalization and is hence exempt from any institutional review.

Results

Overview

Descriptive statistics were tabulated for individual clinical vignettes and aggregated data (Tables 3 and 4). Clinical vignette analysis compared how the readability scores changed with education level for each individual clinical case (DM and HTN). When reported for individual clinical vignettes, data have been reported as AI model-clinical case-education level. Importantly, 2 readability scores were implemented (FKGL and FKRE), so clinical vignette analysis includes a discussion of how both scores change with education level in each individual clinical example.

In this study, accuracy was defined as a readability score (FKRE or FKGL) whose mean, plus or minus one SD, falls within or below the predefined category. For FKRE, these categories are detailed in Table 2, as originally established by Kincaid et al

[18] while for FKGL, the categories are inherently reflected by the corresponding grade levels they represent. For instance, an FKGL score of 6.32 is approximately indicative of a reading level between the 6th and 7th grades. It is important to note that the FKGL formula typically rounds to the nearest whole number, thus for practical purposes, a score of 6.32 is considered appropriate for the 6th grade.

Aggregated data analysis consisted of descriptive statistical reporting similar to the clinical vignette analysis except data were pooled by education level. For example, both of the clinical scenarios were iterated 5 times using “6th-grade” as the prespecified education level. Aggregated data analysis involved pooling the readability scores of all prompt structures that implemented “6th-grade” as the education level (n=10) to observe the consistency of readability scores for educational level across clinical vignettes. FKGL and FKRE scores were acquired, so both metrics were implemented in aggregated data analysis.

Table 3. Mean and SD of FKRE^a and FKGL^b for each education level within each clinical vignette.

AI ^c model	Clinical scenario	Grade level	FKRE, mean (SD)	FKGL, mean (SD)
GPT-4	DM ^d	6th	74.52 (3.12)	6.32 (0.91)
GPT-4	DM	8th	69.42 (3.00)	7.12 (0.91)
GPT-4	DM	HS ^e	47.02 (7.86)	11.4 (1.66)
GPT-4	DM	BS ^f	14.48 (3.33)	16.78 (1.13)
GPT-4	HTN ^g	6th	74.56 (2.34)	6.28 (0.63)
GPT-4	HTN	8th	73.08 (6.17)	6.28 (1.22)
GPT-4	HTN	HS	48.2 (4.69)	10.78 (0.75)
GPT-4	HTN	BS	12.94 (7.89)	17.28 (1.15)
GPT-3.5	DM	6th	54.6 (3.05)	9.7 (0.55)
GPT-3.5	DM	8th	53.6 (6.39)	10.0 (1.19)
GPT-3.5	DM	HS	30.36 (4.81)	13.88 (0.86)
GPT-3.5	DM	BS	26.5 (5.65)	14.44 (1.05)
GPT-3.5	HTN	6th	60.9 (4.08)	8.46 (0.74)
GPT-3.5	HTN	8th	48.96 (2.26)	10.54 (0.98)
GPT-3.5	HTN	HS	34.2 (3.70)	12.92 (0.35)
GPT-3.5	HTN	BS	30.12 (4.63)	13.04 (0.91)

^aFKRE: Flesch reading ease score.

^bFKGL: Flesch-Kincaid grade level.

^cAI: artificial intelligence.

^dDM: diabetes mellitus.

^eHS: high school.

^fBS: bachelor's degree.

^gHTN: hypertension.

Table 4. Descriptive statistics for FKRE^a and FKGL^b scores. All 3 clinical scenarios (diabetes and hypertension) scores are aggregated by education level.

AI ^c model	Grade level	n	FKRE, mean (SD)	FKGL, mean (SD)
GPT-4	6th	10	74.54 (2.6)	6.3 (0.73)
GPT-4	8th	10	71.25 (4.96)	6.7 (1.11)
GPT-4	HS ^d	10	47.61 (6.13)	11.09 (1.26)
GPT-4	BS ^e	10	13.71 (5.77)	17.03 (1.11)
GPT-3.5	6th	10	57.75 (4.75)	9.08 (0.90)
GPT-3.5	8th	10	51.28 (5.14)	10.27 (1.06)
GPT-3.5	HS	10	32.28 (4.52)	13.4 (0.80)
GPT-3.5	BS	10	28.31 (5.22)	13.74 (1.18)

^aFKRE: Flesch reading ease score.

^bFKGL: Flesch-Kincaid grade level.

^cAI: artificial intelligence.

^dHS: high school.

^eBS: bachelor's degree.

Descriptive Statistics—Clinical Vignette Data

Analysis of each group (ie, AI model-clinical case-education level) revealed that GPT-4 consistently produced accurate average FKRE scores for both DM and HTN scenarios across all education levels, with the exception of the 6th grade, where the FKRE scores were 74.52 (SD 3.12) and 74.56 (SD 2.34), respectively (Table 3). Regarding FKGL measures, GPT-4 achieved the target readability for all education levels except for the bachelor's degree for the HTN scenario, where the average FKGL was slightly higher at 17.28 (SD 1.15; Table 3). Conversely, GPT-3.5 accurately produced FKRE and FKGL scores that met the required standards only when tasked with generating outputs for bachelor's degree holders (Table 3). Specifically, in the diabetes scenario at this education level (GPT-3.5-DM-bachelor's degree), FKRE was 26.5 (SD 5.65) and FKGL was 14.44 (SD 1.05), while in the HTN scenario (GPT-3.5-DM-bachelor's degree), FKRE and FKGL scores were 30.12 (SD 4.63) and 13.04 (SD 0.91), respectively (Table 3).

The data from clinical vignettes showed that SDs were stable across subgroups for both GPT-3.5 and GPT-4 (Table 3). The average FKRE SD for GPT-3.5 was 4.91 and for GPT-4 was 4.86 (Table 3). The average FKGL SDs were 0.99 for GPT-3.5 and 1.05 for GPT-4 (Table 3). The highest FKRE SD recorded was 7.89 in the GPT-4 HTN-bachelor's degree scenario, and the lowest was 2.26 in the GPT-3.5 HTN-8th grade scenario (Table 3). For FKGL, the highest SD was 1.66 in the GPT-4 diabetes-high school scenario, and the lowest was 0.35 in the GPT-3.5 HTN-high school scenario (Table 3).

Descriptive Statistics—Aggregated Data

Data were aggregated for each education level across clinical vignettes as mentioned in the Results Overview section. When

aggregated, GPT-4 generated accurate average FKRE scores for most education levels; however, the 6th grade was an exception with an average FKRE of 74.54 (SD 2.60; Table 4). Furthermore, the aggregated data for GPT-4 indicated that the FKGL average was accurate across all tested educational levels (Table 4). In contrast, GPT-3.5 achieved accurate mean FKRE and FKGL scores only at the bachelor's degree level, with averages of 28.31 (SD 5.22) and 13.74 (SD 1.18), respectively (Table 4).

ANOVA and the Tukey Post Hoc Analysis

To determine the differences between the means of each clinical vignette's FKRE and FKGL, unidirectional ANOVA and the Tukey multiple comparison post hoc analysis were performed (Figures 1A-3D). The Tukey post hoc analysis showed significant differences between almost all education levels across both clinical vignettes, both individually and when aggregated (Figures 1A-3D). Notably, in the GPT-4 analysis (both individually and aggregated), the only education levels without a statistically significant difference were between 6th grade and 8th grade, for both FKRE and FKGL (Figures 1A, 1B, 2A, 2B, 3A, and 3B). In the GPT-3.5 DM scenario, this pattern persisted, with an additional absence of significance between the high school and bachelor's education levels for both FKRE and FKGL (Figures 1C, 1D, 2C, 2D, 3C, and 3D). In the GPT-3.5 HTN scenario, the only pair without a significant difference was between high school and bachelor's degree for both FKRE and FKGL (Figures 2C and 2D). Finally, in the aggregated GPT-3.5 data, FKGL showed no significant differences between 6th-grade and 8th-grade or between high school and bachelor's degree, while FKRE lacked significance only between high school and bachelor's degree (Figures 3C and 3D).

Figure 1. (A) GPT-4 diabetes FKRE, compared with single-factor ANOVA and Tukey post hoc test. (B) GPT-4 diabetes FKGL, compared with single-factor ANOVA and Tukey post hoc Test. (C) GPT-3.5 diabetes FKRE, compared with single-factor ANOVA and Tukey post hoc test. (D) GPT-3.5 diabetes FKGL, compared with single-factor ANOVA and Tukey post hoc test. FKRE: Flesch reading ease score; FKGL: Flesch-Kincaid grade level.

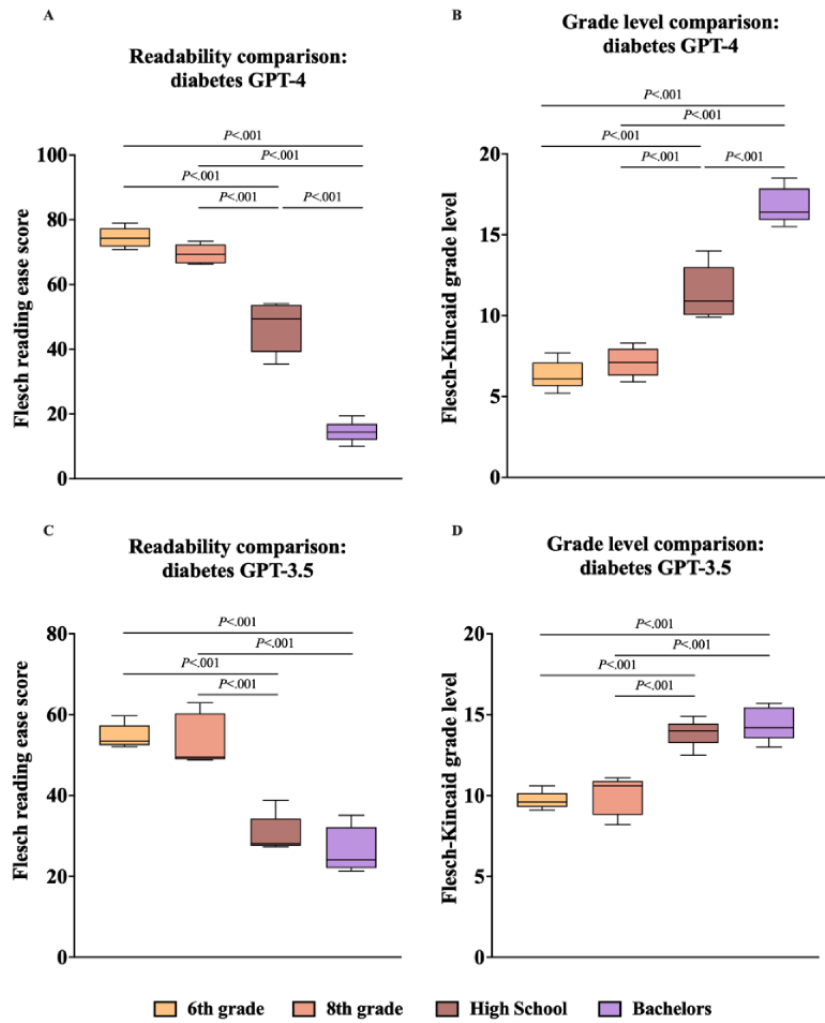


Figure 2. (A) GPT-4 HTN FKRE, compared with single-factor ANOVA and Tukey post hoc test. (B) GPT-4 HTN FKGL, compared with single-factor ANOVA and Tukey post hoc test. (C) GPT-3.5 HTN FKRE, compared with single-factor ANOVA and Tukey post hoc test. (D) GPT-3.5 HTN FKGL, compared with single-factor ANOVA and Tukey post hoc test. FKRE: Flesch reading ease score; FKGL: Flesch-Kincaid grade level; HTN: hypertension.

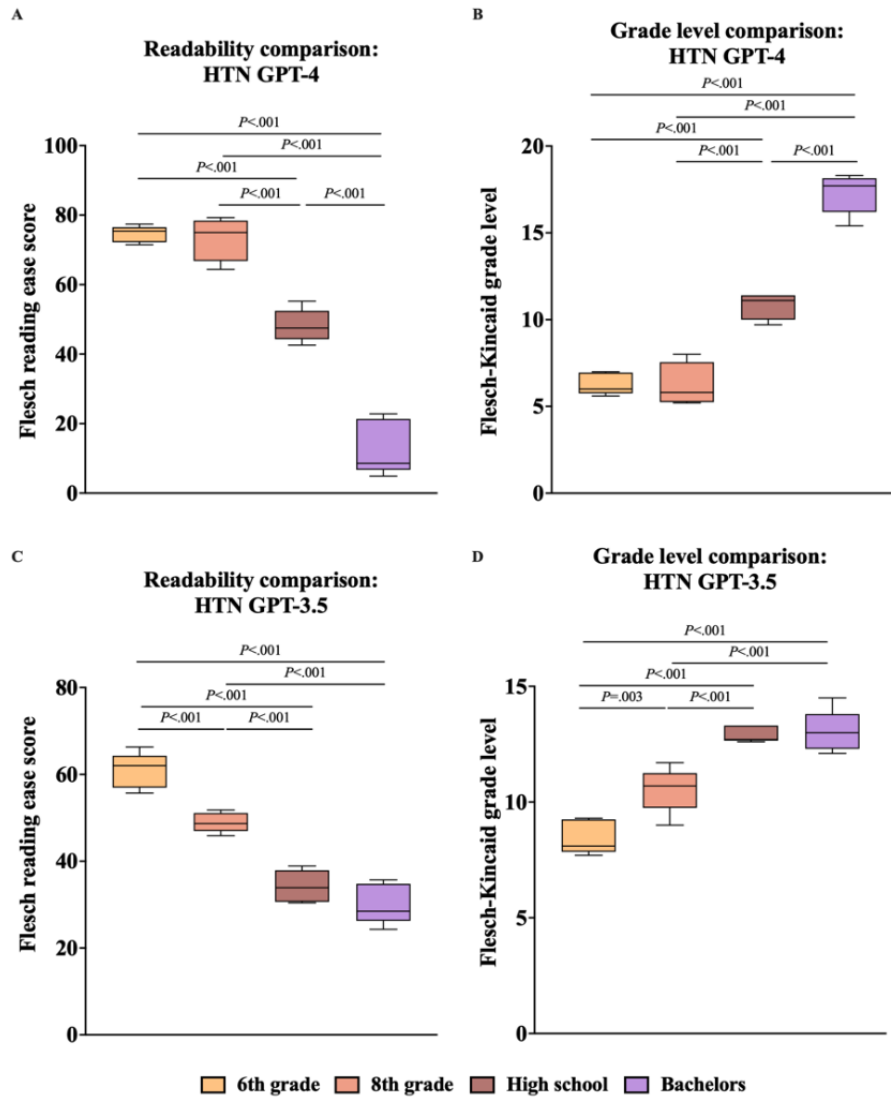
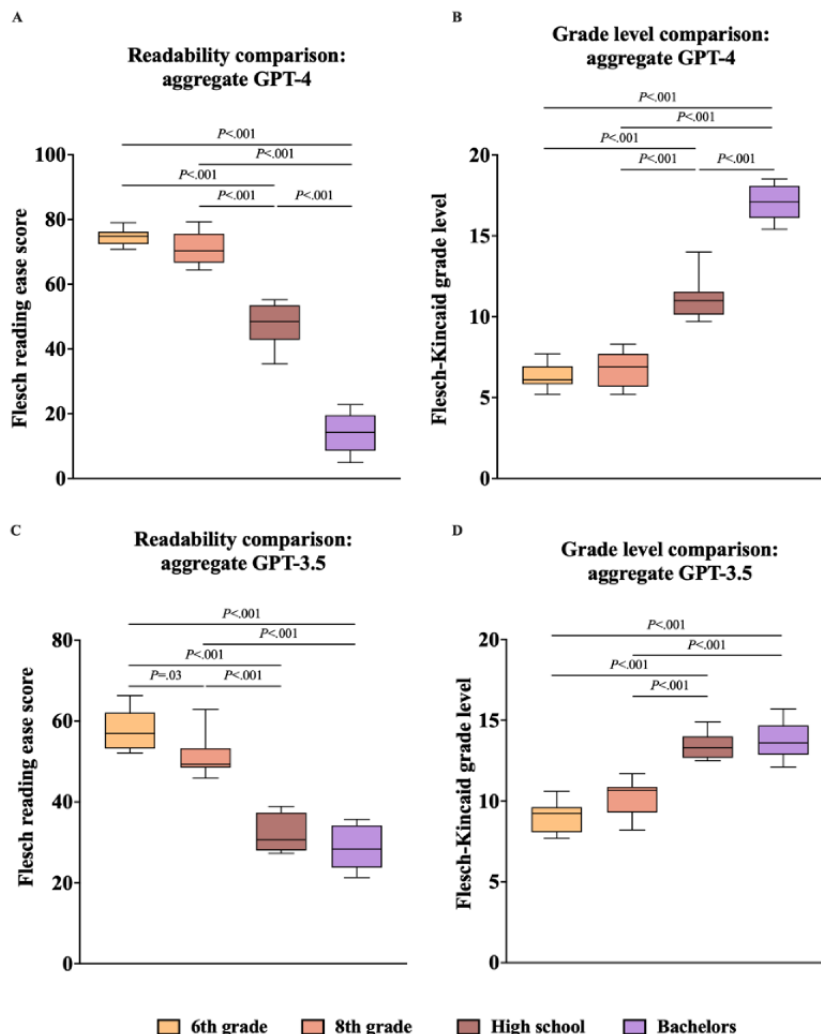


Figure 3. (A) GPT-4 aggregated FKRE, compared with single-factor ANOVA and Tukey post hoc test. (B) GPT-4 aggregated FKGL, compared with single-factor ANOVA and Tukey post hoc test. (C) GPT-3.5 aggregated FKRE, compared with single-factor ANOVA and Tukey post hoc test. (D) GPT-3.5 aggregated FKGL, compared with single-factor ANOVA and Tukey post hoc test. FKRE: Flesch reading ease score; FKGL: Flesch-Kincaid grade level.



Unpaired 2-Tailed *t* Test Analysis—GPT-3.5 Versus GPT-4

When comparing readability scores by education level, unpaired 2-tailed *t* test analysis of individual and aggregated data

consistently showed statistically significant differences between GPT-4 and GPT-3.5 (Figures 4A-6H). The analysis revealed that GPT-4 generally produced more readable outputs (higher FKRE and lower FKGL) across all education levels, except for the bachelor’s degree (Figures 4A-6H).

Figure 4. (A) Comparison of FKRE between GPT-4 and GPT-3.5 for diabetes outputs at the 6th-grade level, analyzed with an unpaired 2-tailed *t* test. (B) Comparison of FKGL between GPT-4 and GPT-3.5 for diabetes outputs at the 6th-grade level, analyzed with an unpaired 2-tailed *t* test. (C) Comparison of FKRE between GPT-4 and GPT-3.5 for diabetes outputs at the 8th-grade level, analyzed with an unpaired 2-tailed *t* test. (D) Comparison of FKGL between GPT-4 and GPT-3.5 for diabetes outputs at the 8th-grade level, analyzed with an unpaired 2-tailed *t* test. (E) Comparison of FKRE between GPT-4 and GPT-3.5 for diabetes outputs at the high school level, analyzed with an unpaired 2-tailed *t* test. (F) Comparison of FKGL between GPT-4 and GPT-3.5 for diabetes outputs at the high school level, analyzed with an unpaired 2-tailed *t* test. (G) Comparison of FKRE between GPT-4 and GPT-3.5 for diabetes outputs at the bachelor's level, analyzed with an unpaired 2-tailed *t* test. (H) Comparison of FKGL between GPT-4 and GPT-3.5 for diabetes outputs at the bachelor's level, analyzed with an unpaired 2-tailed *t* test. FKRE: Flesch reading ease score; FKGL: Flesch-Kincaid grade level.

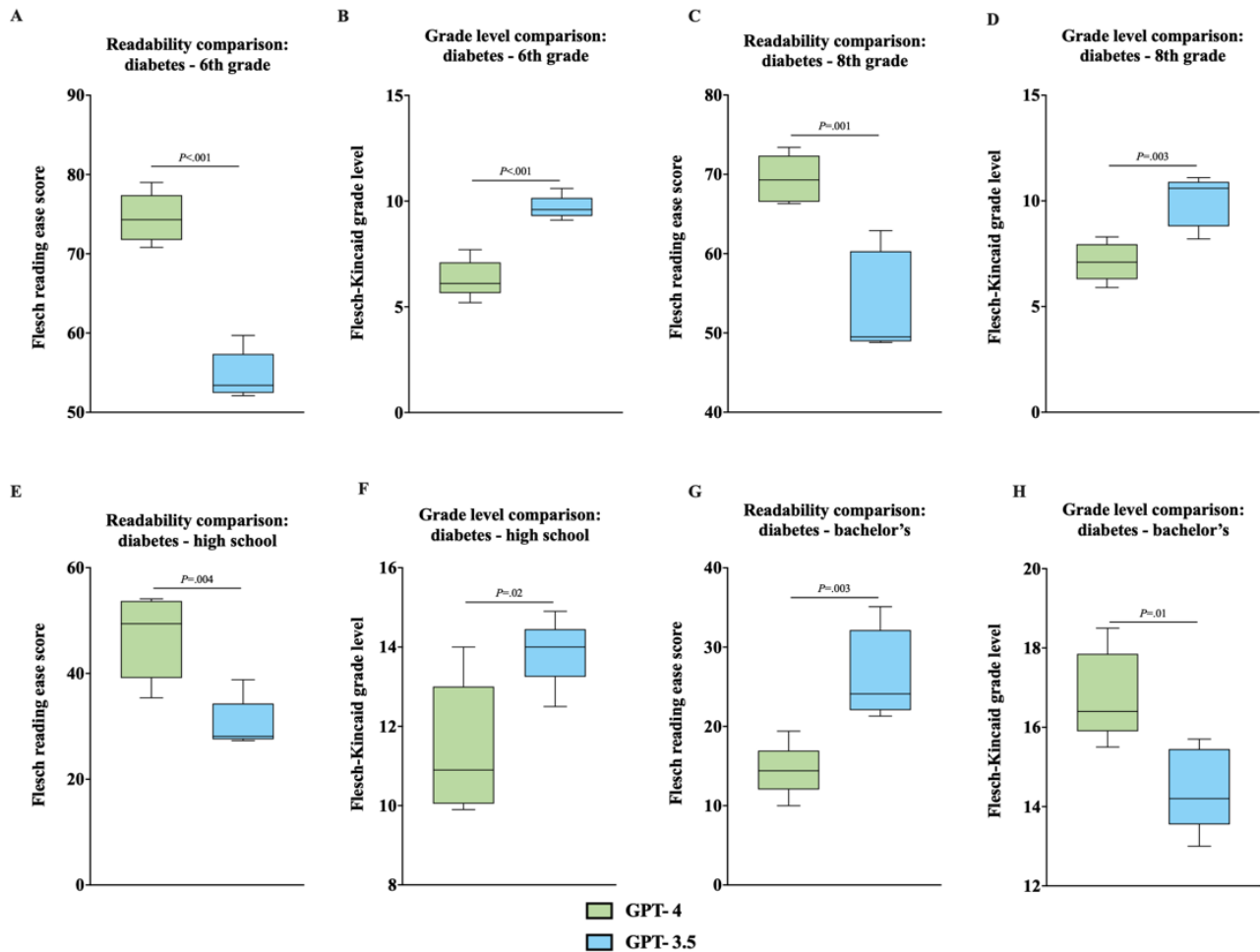


Figure 5. (A) Comparison of FKRE between GPT-4 and GPT-3.5 for HTN outputs at the 6th-grade level, analyzed with an unpaired 2-tailed *t* test. (B) Comparison of FKGL between GPT-4 and GPT-3.5 for HTN outputs at the 6th-grade level, analyzed with an unpaired 2-tailed *t* test. (C) Comparison of FKRE between GPT-4 and GPT-3.5 for HTN outputs at the 8th-grade level, analyzed with an unpaired 2-tailed *t* test. (D) Comparison of FKGL between GPT-4 and GPT-3.5 for HTN outputs at the 8th-grade level, analyzed with an unpaired 2-tailed *t* test. (E) Comparison of FKRE between GPT-4 and GPT-3.5 for HTN outputs at the high school level, analyzed with an unpaired 2-tailed *t* test. (F) Comparison of FKGL between GPT-4 and GPT-3.5 for HTN outputs at the high school level, analyzed with an unpaired 2-tailed *t* test. (G) Comparison of FKRE between GPT-4 and GPT-3.5 for HTN outputs at the bachelor's level, analyzed with an unpaired 2-tailed *t* test. (H) Comparison of FKGL between GPT-4 and GPT-3.5 for HTN outputs at the bachelor's level, analyzed with an unpaired 2-tailed *t* test. FKRE: Flesch reading ease score; FKGL: Flesch-Kincaid grade level; HTN: hypertension.

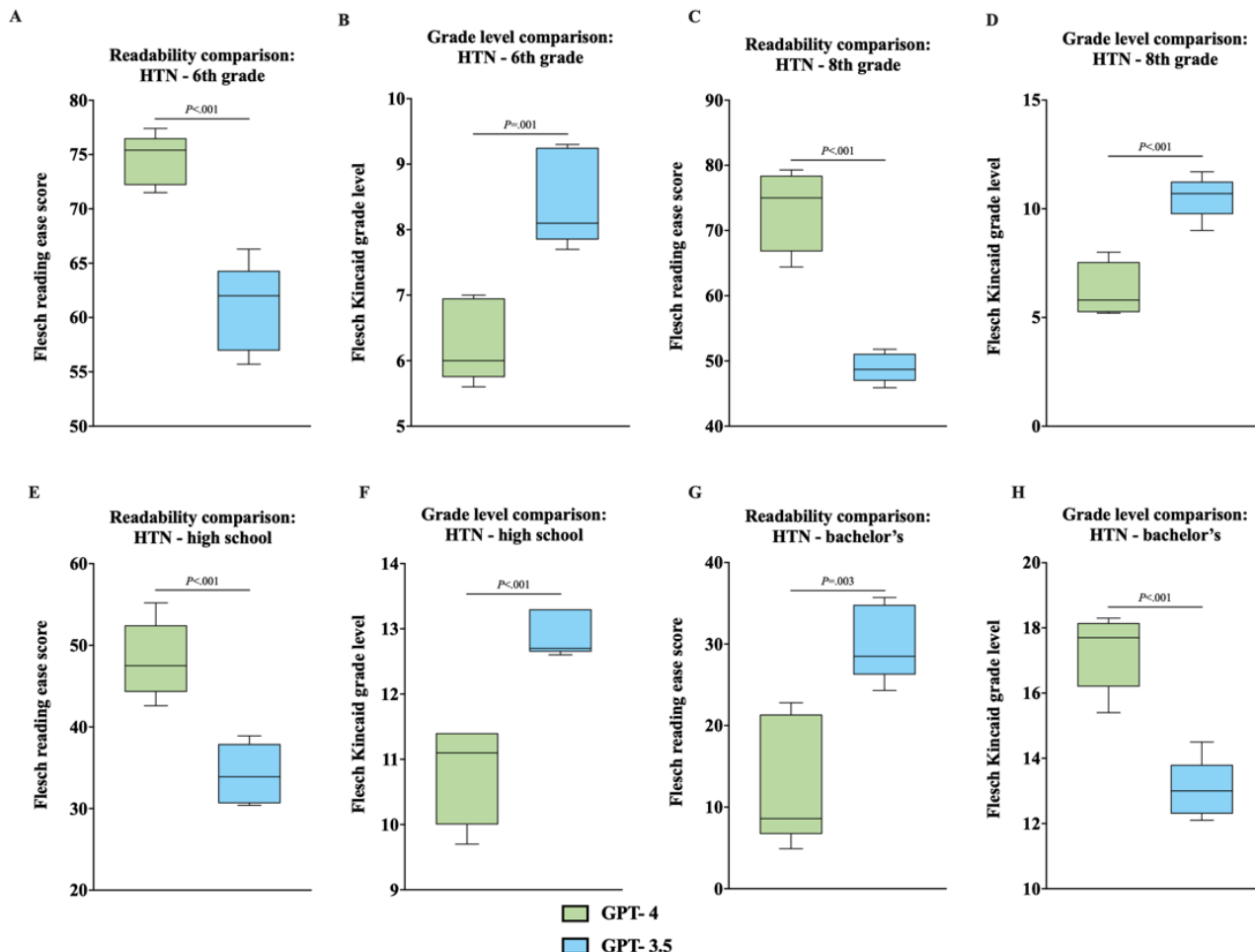
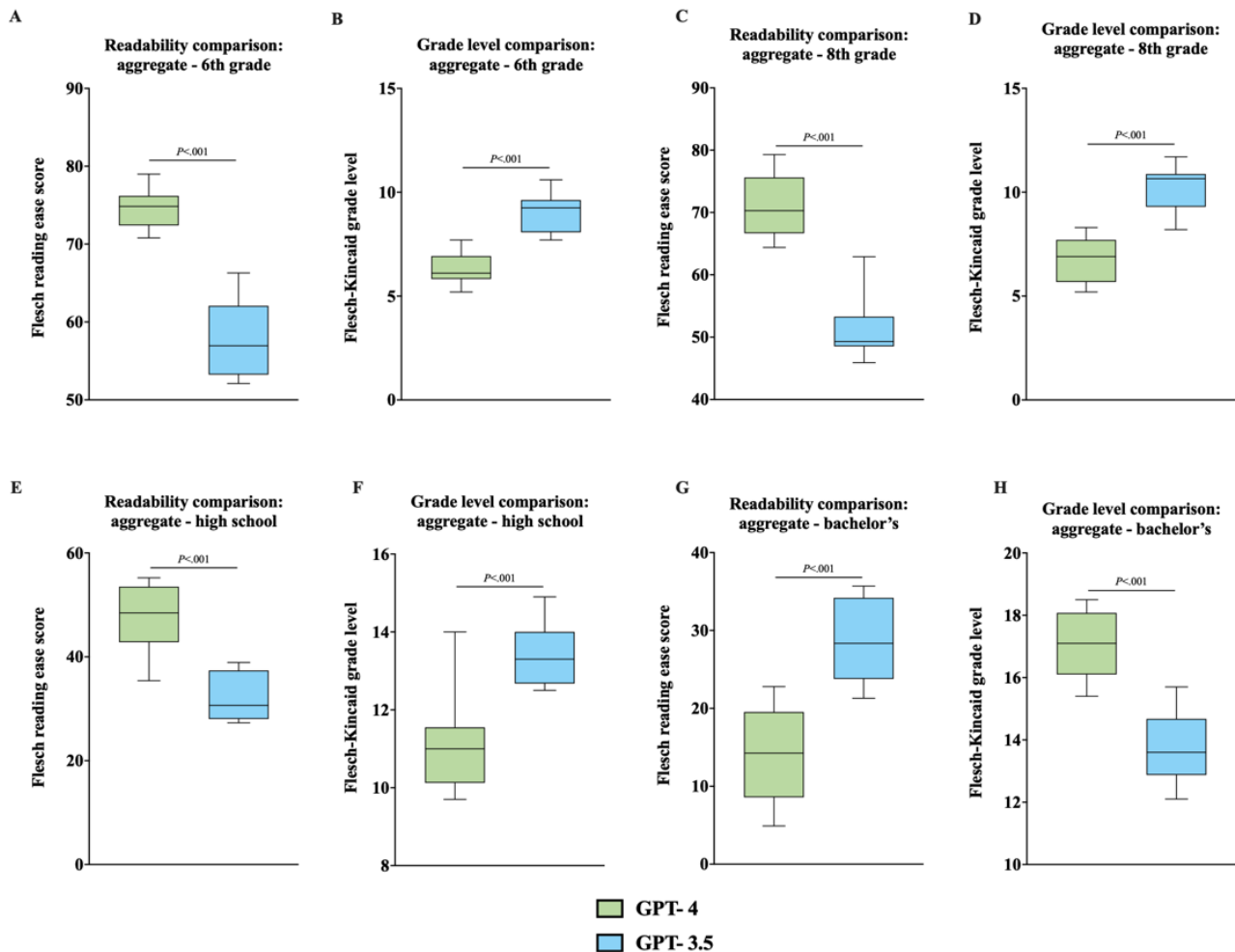


Figure 6. (A) Comparison of FKRE between GPT-4 and GPT-3.5 for aggregated outputs at the 6th-grade level, analyzed with an unpaired 2-tailed *t* test. (B) Comparison of FKGL between GPT-4 and GPT-3.5 for aggregated outputs at the 6th-grade level, analyzed with an unpaired 2-tailed *t* test. (C) Comparison of FKRE between GPT-4 and GPT-3.5 for aggregated outputs at the 8th-grade level, analyzed with an unpaired 2-tailed *t* test. (D) Comparison of FKGL between GPT-4 and GPT-3.5 for aggregated outputs at the 8th-grade level, analyzed with an unpaired 2-tailed *t* test. (E) Comparison of FKRE between GPT-4 and GPT-3.5 for aggregated outputs at the high school level, analyzed with an unpaired 2-tailed *t* test. (F) Comparison of FKGL between GPT-4 and GPT-3.5 for aggregated outputs at the high school level, analyzed with an unpaired 2-tailed *t* test. (G) Comparison of FKRE between GPT-4 and GPT-3.5 for aggregated outputs at the bachelor's level, analyzed with an unpaired 2-tailed *t* test. (H) Comparison of FKGL between GPT-4 and GPT-3.5 for aggregated outputs at the bachelor's level, analyzed with an unpaired 2-tailed *t* test. FKRE: Flesch reading ease score; FKGL: Flesch-Kincaid grade level.



Discussion

Overview

Previous investigations into the use of ChatGPT within health care primarily focused on evaluating its potential as an educational tool for patients, particularly in terms of content accuracy and the general readability of its outputs [15,28,29]. Previous studies have also explored the capacity of ChatGPT to distill complex medical information, such as published research abstracts, thereby enhancing accessibility for patients lacking specialized medical knowledge [16,30]. Despite these advancements, no research to date has specifically investigated ChatGPT's ability to adjust the readability of its outputs to match different educational levels as explicitly directed by users. This study aimed to fill that gap by assessing the capacity of GLMs to produce tailored educational content that adheres to specified readability standards based on user input.

Principal Results

Analysis of the FKRE data showed some trends that point to GPT-4 having the potential to achieve this goal (Figures 1A-3D). GPT-4 can consistently generate outputs at 3 generalized reading levels: easy (6th and 8th grade), medium (high school), and difficult (bachelor's degree). In the case of GPT-4, the readability analysis revealed indistinct results exclusively between the outputs for the 6th and 8th grades (Figures 1A, 1B, 2A, 2B, 3A, and 3B). This indistinguishability likely stems from the close progression of these 2 educational stages, being solely separated by the 7th grade. In contrast, all other adjacent educational levels examined in this study were separated by a minimum of 4 grades, which inherently facilitated a more distinct comparison. Although the differences in readability between the outputs for the 6th and 8th grades were not statistically significant, further investigation is warranted to ascertain whether this similarity has any substantive implications for clinical or educational outcomes when these outputs are used in patient education.

Similar to GPT-4, GPT-3.5 also demonstrated nonsignificant differences in readability between the 6th- and 8th-grade levels, both in individual DM scenarios and when considering the overall FKGL. Although the difference in readability scores for the HTN scenario and the combined FKGL scenario was statistically significant between these grades, the scores were higher than the target readability level across the board (Tables 3 and 4 and Figures 2C, 2D, and 3C). One distinct deviation in performance between GPT-3.5 and GPT-4 was the former's consistent failure to produce outputs with a significant difference in readability between the high school and bachelor's degree levels across all test cases (Figures 1C, 1D, 2C, 2D, 3C, and 3D). This suggests that GPT-3.5 may be less adept than GPT-4 at differentiating between education levels in its generated text when given specific prompts.

Finally, for the mean FKRE, the trend between education levels was always negative, meaning that as the education level increased, the prompts became harder to read (Figures 1A, 1C, 2A, 2C, 3A, and 3C). Analysis of the FKGL data also showed similarly consistent trends as FKGL average scores always increased with higher education levels (Figures 1B, 1D, 2B, 2D, 3B, and 3D). This is encouraging, as these results show even at this early stage of its existence, GLMs, such as GPT-3.5 and GPT-4, can consistently create outputs of varying readability when explicitly prompted by an input.

GPT-3.5 and GPT-4 demonstrated relatively consistent results in variability across all educational levels, suggesting that both versions of ChatGPT maintain uniform performance irrespective of the complexity of language in the clinical vignettes used. Repeated trials across scenarios—conducted 5 times each—affirmed the reliability of our findings, as consistency in outputs was systematically verified. Notably, the average SDs for the FKRE scores were 4.91 for GPT-3.5 and 4.86 for GPT-4, respectively. Given that these values are less than 5 and considering that a 10-point difference on the FKRE scale roughly corresponds to one grade level (as detailed in Table 1), it can be inferred that 95% of the FKRE scores for both models are expected to cluster within one grade level of each other. This is significant as it highlights the models' ability to produce outputs with stable readability values, with most variations not deviating dramatically from the mean. Similarly, the average FKGL SDs were 0.99 for GPT-3.5 and 1.05 for GPT-4, indicating that roughly 95% of FKGL scores likely fall within approximately two grade levels, providing further evidence of output consistency. It is important to clarify that this analysis does not assess the accuracy of the outputs in matching the requested readability levels but rather their consistency in reaching said levels.

A key difference in performance between GPT-3.5 and GPT-4 was observed in the accuracy of the outputs' readability levels. GPT-4 achieved accurate average readability scores in 13 out of 16 scenarios across both FKRE and FKGL, while GPT-3.5 reached accurate average readability scores in only 4 out of 16 scenarios, exclusively at the bachelor's degree education level (Tables 2 and 3). A comparative grade level analysis using an unpaired 2-tailed *t* test, both for individual and aggregated data, consistently indicated statistically significant differences between GPT-4 and GPT-3.5. This analysis suggests that GPT-4

generally delivered outputs with better readability (higher FKRE and lower FKGL) across various educational levels, with the exception of the bachelor's degree scenarios (Figures 4A-6H). These findings validate our hypothesis that GPT-4 would outperform its predecessor in output readability accuracy, highlighting its improved language processing capabilities. This suggests that GPT-4 could be more effective in applications requiring nuanced understanding and generation of text such as educational tools or automated content creation. Future research could explore the specific enhancements in GPT-4 that contribute to these improvements and test its performance in other domains to further understand its broader applicability and limitations.

FKRE and FKGL scores were implemented, as they weigh aspects of readability differently (equations 1 and 2). The FKGL emphasizes sentence length more than word length when compared to FKRE [18]. This explains some of the inconsistency in the trend analysis of group variance. Ultimately, our findings concerning FKRE and FKGL scores examine GPT-3.5 and GPT-4's ability to reliably respond to varying education levels, which as a clinical tool, has the potential to be beneficial in educating patients [31]. However, future research must quantify readability with more metrics to ensure proper personalization of patient-facing educational information.

Our results indicate that GLMs have the potential to create customizable educational materials for patients, suggesting a possible role as a new tool in addressing LHL. Further research is integral in elucidating the capacity that ChatGPT and other GLMs can address LHL, as a patient's level of health literacy can significantly impact their health outcomes [32]. Specifically, patients with LHL have higher hospitalization rates, are more likely to have poor health status, and have a mortality rate almost double that of patients who do not have LHL [3]. These patients are less likely to receive preventive health services and are more likely to face difficulty accessing the health care they require [33,34]. Current services addressing LHL, including educational pamphlets and community health fairs, have shown limited success due to accessibility constraints [4,35,36]. Thus, attempts to bridge this gap in health literacy and improve health outcomes have been focused on improving health communication techniques for patients with LHL [3]. In this regard, ChatGPT and other new technologies exhibit clear potential, however, are not currently suitable for clinical use in this context. Use of either GPT-3.5 or GPT-4 is not recommended with patients, at the time of this publication due to a few significant limitations.

Limitations

The major limitation of this study was that it did not analyze the accuracy of the content produced by ChatGPT. Other studies have elucidated the accuracy of ChatGPT outputs in the context of patient queries, particularly within the fields of otolaryngology, urology, and plastic surgery. These studies demonstrated that while ChatGPT can provide accurate answers to patient-style questions, it often answers questions incorrectly [37-39]. Without thorough content validation, the use of GLM technology to generate patient education materials could inadvertently contain false information, potentially leading to the spread of misinformation and resulting in patients

mishandling their own care. At present, we strongly recommend that patients seeking medical information obtain their health education materials directly from a qualified physician. Specifically, the use of natural language processors, like ChatGPT, should be restricted to licensed health care providers when disseminating information to patients. This approach ensures the accuracy and reliability of the health information provided. Additionally, this study only tested 2 clinical scenarios. These scenarios centered around a patient using ChatGPT to learn about a new diagnosis of DM or HTN. This study design potentially limits the generalizability of these findings in other clinical contexts.

Another challenge GLMs face is related to maintaining patient privacy [11]. In March of 2023, OpenAI confirmed that they experienced a data leak in which select conversation titles from random users were made visible to other users [11]. The comprehensive impact of this data breach is unclear, but the prospect of future breaches—particularly those involving protected health information—represents a substantial privacy concern for GLMs [11]. At this time, there is no way for a publicly accessible GLM, such as ChatGPT, to be trained on protected health information while maintaining Health Insurance Portability and Accountability Act compliance [40]. Companies and health care professionals looking to develop GLMs for medical use continue to face legal hurdles aimed at protecting

patient privacy [41]. In addition to data leaks, ChatGPT was functionally banned for a week in Italy in March of 2023 due to accusations that it was violating European Union data protection laws [42]. This ban prompted other countries, including Germany, Spain, and Canada, to launch investigations into ChatGPT [42]. Given these valid concerns regarding privacy and legality, developers should continue to address these challenges as they integrate this technology into medical care.

Conclusions

GPT-4 can create outputs within 3 tiers of readability: easy (6th and 8th grade), medium (high school), and difficult (bachelor's degree). These 3 tiers fall relatively well into their correct intended levels of readability according to the FKRE and FKGL and they allow for preliminary stratification of readability. Unfortunately, GPT-3.5 is less adept at creating customized outputs that fall into their specified readability ranges. Our results highlight GPT-4's ability to provide patient-centered responses with statistically significant changes to output readability based on education level. Further optimization of this personalization's accuracy is necessary for it to be an effective clinical tool in addressing LHL. This must be coupled with comprehensive content validation and stringent privacy security measures. The continued evolution of GLMs should provide more robust and capable tools to address these limitations in order to best educate and empower patients.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Revision.

[[DOCX File , 2849 KB - ai_v3i1e54371_app1.docx](#)]

References

1. Kutner MGE, Jin Y, Paulsen C. The health literacy of America's adults: results from the 2003 National Assessment of Adult Literacy. NCES 2006-483. Washington, DC: National Center for Education Statistics; 2006. URL: <https://nces.ed.gov/pubs2006/2006483.pdf> [accessed 2024-07-19]
2. Institute of Medicine, Board on Neuroscience and Behavioral Health. In: Nielsen-Bohlman L, Panzer AM, Kindig DA, editors. Health Literacy: A Prescription to End Confusion. Washington DC: National Academies Press; 2004.
3. Sudore RL, Schillinger D. Interventions to improve care for patients with limited health literacy. *J Clin Outcomes Manag* 2009;16(1):20-29 [FREE Full text] [Medline: 20046798]
4. Murray K, Liang A, Barnack-Tavlaris J, Navarro AM. The reach and rationale for community health fairs. *J Cancer Educ* 2014 Mar;29(1):19-24 [FREE Full text] [doi: 10.1007/s13187-013-0528-3] [Medline: 23907787]
5. Internet, broadband fact sheet. Washington, DC: Pew Research Center; 2024. URL: <https://www.pewresearch.org/internet/fact-sheet/internet-broadband/> [accessed 2024-07-19]
6. Lee K, Hoti K, Hughes JD, Emmerton L. Dr Google and the consumer: a qualitative study exploring the navigational needs and online health information-seeking behaviors of consumers with chronic health conditions. *J Med Internet Res* 2014 Dec 02;16(12):e262 [FREE Full text] [doi: 10.2196/jmir.3706] [Medline: 25470306]
7. Yeo YH, Samaan JS, Ng WH, Ting PS, Trivedi H, Vipani A, et al. Assessing the performance of ChatGPT in answering questions regarding cirrhosis and hepatocellular carcinoma. *Clin Mol Hepatol* 2023 Jul;29(3):721-732 [FREE Full text] [doi: 10.3350/cmh.2023.0089] [Medline: 36946005]
8. Zemčík MT. A brief history of chatbots. : DEStech Transactions on Computer Science and Engineering; 2019 Presented at: International Conference on Artificial Intelligence, Control and Automation Engineering (AICAE 2019); June 23-24, 2019; Wuhan, China URL: https://www.researchgate.net/profile/Tomas-Zemcik/publication/336734161_A_Brief_History_of_Chatbots/links/5dc1bc51a6fdcc21280872a3/A-Brief-History-of-Chatbots.pdf/>

9. Scherr R, Halaseh FF, Spina A, Andalib S, Rivera R. ChatGPT interactive medical simulations for early clinical education: case study. *JMIR Med Educ* 2023 Nov 10;9:e49877 [FREE Full text] [doi: [10.2196/49877](https://doi.org/10.2196/49877)] [Medline: [37948112](https://pubmed.ncbi.nlm.nih.gov/37948112/)]
10. Jayakumar P, Moore MG, Furlough KA, Uhler LM, Andrawis JP, Koenig KM, et al. Comparison of an artificial intelligence-enabled patient decision aid vs educational material on decision quality, shared decision-making, patient experience, and functional outcomes in adults with knee osteoarthritis: a randomized clinical trial. *JAMA Netw Open* 2021 Feb 01;4(2):e2037107 [FREE Full text] [doi: [10.1001/jamanetworkopen.2020.37107](https://doi.org/10.1001/jamanetworkopen.2020.37107)] [Medline: [33599773](https://pubmed.ncbi.nlm.nih.gov/33599773/)]
11. Sallam M. ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. *Healthcare (Basel)* 2023 Mar 19;11(6):887 [FREE Full text] [doi: [10.3390/healthcare11060887](https://doi.org/10.3390/healthcare11060887)] [Medline: [36981544](https://pubmed.ncbi.nlm.nih.gov/36981544/)]
12. Sun F, Zimmer Z, Zajacova A. Pain and disability transitions among older Americans: the role of education. *J Pain* 2023 Jun;24(6):1009-1019 [FREE Full text] [doi: [10.1016/j.jpain.2023.01.014](https://doi.org/10.1016/j.jpain.2023.01.014)] [Medline: [36706888](https://pubmed.ncbi.nlm.nih.gov/36706888/)]
13. Huh S. Can we trust AI chatbots' answers about disease diagnosis and patient care? *J Korean Med Assoc* 2023 Apr;66(4):218-222. [doi: [10.5124/jkma.2023.66.4.218](https://doi.org/10.5124/jkma.2023.66.4.218)]
14. Baumann J, Marshall S, Groneck A, Hanish SJ, Choma T, DeFroda S. Readability of spine-related patient education materials: a standard method for improvement. *Eur Spine J* 2023 Sep;32(9):3039-3046. [doi: [10.1007/s00586-023-07856-5](https://doi.org/10.1007/s00586-023-07856-5)] [Medline: [37466719](https://pubmed.ncbi.nlm.nih.gov/37466719/)]
15. Davis R, Eppler M, Ayo-Ajibola O, Loh-Doyle JC, Nabhani J, Samplaski M, et al. Evaluating the effectiveness of artificial intelligence-powered large language models application in disseminating appropriate and readable health information in urology. *J Urol* 2023 Oct;210(4):688-694. [doi: [10.1097/JU.0000000000003615](https://doi.org/10.1097/JU.0000000000003615)] [Medline: [37428117](https://pubmed.ncbi.nlm.nih.gov/37428117/)]
16. Eppler MB, Ganjavi C, Knudsen JE, Davis RJ, Ayo-Ajibola O, Desai A, et al. Bridging the gap between urological research and patient understanding: the role of large language models in automated generation of layperson's summaries. *Urol Pract* 2023 Sep;10(5):436-443. [doi: [10.1097/UPJ.0000000000000428](https://doi.org/10.1097/UPJ.0000000000000428)] [Medline: [37410015](https://pubmed.ncbi.nlm.nih.gov/37410015/)]
17. Kirchner GJ, Kim RY, Weddle JB, Bible JE. Can artificial intelligence improve the readability of patient education materials? *Clin Orthop Relat Res* 2023 Nov 01;481(11):2260-2267. [doi: [10.1097/CORR.0000000000002668](https://doi.org/10.1097/CORR.0000000000002668)] [Medline: [37116006](https://pubmed.ncbi.nlm.nih.gov/37116006/)]
18. Kincaid JP, Fishburne RP, Rogers RL, Chissom BS. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Institute for simulation and training. 1975. URL: <https://stars.library.ucf.edu/istlibrary/56/> [accessed 2024-07-19]
19. Si L, Callan J. A statistical model for scientific readability. 2001 Presented at: Proceedings of the 10th International Conference on Information and Knowledge Management; October 5, 2001; Georgia, Atlanta, USA p. 574-576. [doi: [10.1145/502585.502695](https://doi.org/10.1145/502585.502695)]
20. Badarudeen S, Sabharwal S. Assessing readability of patient education materials: current role in orthopaedics. *Clin Orthop Relat Res* 2010 Oct;468(10):2572-2580 [FREE Full text] [doi: [10.1007/s11999-010-1380-y](https://doi.org/10.1007/s11999-010-1380-y)] [Medline: [20496023](https://pubmed.ncbi.nlm.nih.gov/20496023/)]
21. Williamson JML, Martin AG. Analysis of patient information leaflets provided by a district general hospital by the Flesch and Flesch-Kincaid method. *Int J Clin Pract* 2010 Dec;64(13):1824-1831. [doi: [10.1111/j.1742-1241.2010.02408.x](https://doi.org/10.1111/j.1742-1241.2010.02408.x)] [Medline: [21070533](https://pubmed.ncbi.nlm.nih.gov/21070533/)]
22. Beauchamp A, Buchbinder R, Dodson S, Batterham RW, Elsworth GR, McPhee C, et al. Distribution of health literacy strengths and weaknesses across socio-demographic groups: a cross-sectional survey using the Health Literacy Questionnaire (HLQ). *BMC Public Health* 2015 Jul 21;15:678 [FREE Full text] [doi: [10.1186/s12889-015-2056-z](https://doi.org/10.1186/s12889-015-2056-z)] [Medline: [26194350](https://pubmed.ncbi.nlm.nih.gov/26194350/)]
23. Jansen T, Rademakers J, Waverijn G, Verheij R, Osborne R, Heijmans M. The role of health literacy in explaining the association between educational attainment and the use of out-of-hours primary care services in chronically ill people: a survey study. *BMC Health Serv Res* 2018 May 31;18(1):394 [FREE Full text] [doi: [10.1186/s12913-018-3197-4](https://doi.org/10.1186/s12913-018-3197-4)] [Medline: [29855365](https://pubmed.ncbi.nlm.nih.gov/29855365/)]
24. Safeer RS, Keenan J. Health literacy: the gap between physicians and patients. *Am Fam Physician* 2005;72(3):463-468 [FREE Full text] [Medline: [16100861](https://pubmed.ncbi.nlm.nih.gov/16100861/)]
25. Weiss BD, Schwartzberg JG, American Medical A. Health literacy and patient safety: help patients understand: manual for clinicians. Chicago, Illinois: AMA Foundation; 2007. URL: http://www.hhvna.com/files/Courses/HealthLiteracy/Health_Literacy_Manual_AMA_Revised.pdf [accessed 2024-07-19]
26. Centers for Disease Control and Prevention (U.S.), Office of the Associate Director for Communication, Strategic and Proactive Communication Branch. Simply Put: a guide for creating easy-to-understand materials. CDC 2010:1-44 [FREE Full text]
27. Kelly NEW, Murray KE, McCarthy C, O'Shea DB. An objective analysis of quality and readability of online information on COVID-19. *Health Technol (Berl)* 2021;11(5):1093-1099 [FREE Full text] [doi: [10.1007/s12553-021-00574-2](https://doi.org/10.1007/s12553-021-00574-2)] [Medline: [34189011](https://pubmed.ncbi.nlm.nih.gov/34189011/)]
28. Cappellani F, Card KR, Shields CL, Pulido JS, Haller JA. Reliability and accuracy of artificial intelligence ChatGPT in providing information on ophthalmic diseases and management to patients. *Eye (Lond)* 2024 May;38(7):1368-1373 [FREE Full text] [doi: [10.1038/s41433-023-02906-0](https://doi.org/10.1038/s41433-023-02906-0)] [Medline: [38245622](https://pubmed.ncbi.nlm.nih.gov/38245622/)]
29. Gordon EB, Towbin AJ, Wingrove P, Shafique U, Haas B, Kitts AB, et al. Enhancing patient communication with Chat-GPT in radiology: evaluating the efficacy and readability of answers to common imaging-related questions. *J Am Coll Radiol* 2024 Feb;21(2):353-359. [doi: [10.1016/j.jacr.2023.09.011](https://doi.org/10.1016/j.jacr.2023.09.011)] [Medline: [37863153](https://pubmed.ncbi.nlm.nih.gov/37863153/)]

30. Eid K, Eid A, Wang D, Raiker RS, Chen S, Nguyen J. Optimizing ophthalmology patient education via chatbot-generated materials: readability analysis of AI-generated patient education materials and the American Society of Ophthalmic Plastic and Reconstructive Surgery Patient Brochures. *Ophthalmic Plast Reconstr Surg* 2024;40(2):212-216. [doi: [10.1097/IOP.0000000000002549](https://doi.org/10.1097/IOP.0000000000002549)] [Medline: [37972974](https://pubmed.ncbi.nlm.nih.gov/37972974/)]
31. Rooney MK, Santiago G, Perni S, Horowitz DP, McCall AR, Einstein AJ, et al. Readability of patient education materials from high-impact medical journals: a 20-year analysis. *J Patient Exp* 2021;8:2374373521998847 [FREE Full text] [doi: [10.1177/2374373521998847](https://doi.org/10.1177/2374373521998847)] [Medline: [34179407](https://pubmed.ncbi.nlm.nih.gov/34179407/)]
32. Ferguson LA, Pawlak R. Health literacy: the road to improved health outcomes. *J Nurse Pract* 2011 Feb;7(2):123-129. [doi: [10.1016/j.nurpra.2010.11.020](https://doi.org/10.1016/j.nurpra.2010.11.020)]
33. Baker DW, Gazmararian JA, Williams MV, Scott T, Parker RM, Green D, et al. Health literacy and use of outpatient physician services by Medicare managed care enrollees. *J Gen Intern Med* 2004 Mar;19(3):215-220 [FREE Full text] [doi: [10.1111/j.1525-1497.2004.21130.x](https://doi.org/10.1111/j.1525-1497.2004.21130.x)] [Medline: [15009775](https://pubmed.ncbi.nlm.nih.gov/15009775/)]
34. Sudore RL, Mehta KM, Simonsick EM, Harris TB, Newman AB, Satterfield S, et al. Limited literacy in older people and disparities in health and healthcare access. *J Am Geriatr Soc* 2006 May;54(5):770-776. [doi: [10.1111/j.1532-5415.2006.00691.x](https://doi.org/10.1111/j.1532-5415.2006.00691.x)] [Medline: [16696742](https://pubmed.ncbi.nlm.nih.gov/16696742/)]
35. Bodenheimer T, Pham HH. Primary care: current problems and proposed solutions. *Health Aff (Millwood)* 2010 May;29(5):799-805. [doi: [10.1377/hlthaff.2010.0026](https://doi.org/10.1377/hlthaff.2010.0026)] [Medline: [20439864](https://pubmed.ncbi.nlm.nih.gov/20439864/)]
36. Wolfe MK, McDonald NC, Holmes GM. Transportation barriers to health care in the United States: findings from the National Health Interview Survey, 1997-2017. *Am J Public Health* 2020 Jun;110(6):815-822. [doi: [10.2105/AJPH.2020.305579](https://doi.org/10.2105/AJPH.2020.305579)] [Medline: [32298170](https://pubmed.ncbi.nlm.nih.gov/32298170/)]
37. Maksimoski M, Noble AR, Smith DF. Does ChatGPT answer otolaryngology questions accurately? *Laryngoscope* 2024 Mar 28. [doi: [10.1002/lary.31410](https://doi.org/10.1002/lary.31410)] [Medline: [38545679](https://pubmed.ncbi.nlm.nih.gov/38545679/)]
38. Coskun B, Ocakoglu G, Yetemen M, Kaygisiz O. Can ChatGPT, an artificial intelligence language model, provide accurate and high-quality patient information on prostate cancer? *Urology* 2023 Oct;180:35-58. [doi: [10.1016/j.urology.2023.05.040](https://doi.org/10.1016/j.urology.2023.05.040)] [Medline: [37406864](https://pubmed.ncbi.nlm.nih.gov/37406864/)]
39. Copeland-Halperin LR, O'Brien L, Copeland M. Evaluation of artificial intelligence-generated responses to common plastic surgery questions. *Plast Reconstr Surg Glob Open* 2023 Aug;11(8):e5226 [FREE Full text] [doi: [10.1097/GOX.0000000000005226](https://doi.org/10.1097/GOX.0000000000005226)] [Medline: [37654681](https://pubmed.ncbi.nlm.nih.gov/37654681/)]
40. AI rising: ChatGPT, healthcare, and HIPAA compliance. Compliancy Group. 2023. URL: <https://compliancy-group.com/hipaa-and-chatgpt/> [accessed 2024-07-18]
41. Moor M, Banerjee O, Abad ZSH, Krumholz HM, Leskovec J, Topol EJ, et al. Foundation models for generalist medical artificial intelligence. *Nature* 2023 Apr;616(7956):259-265. [doi: [10.1038/s41586-023-05881-4](https://doi.org/10.1038/s41586-023-05881-4)] [Medline: [37045921](https://pubmed.ncbi.nlm.nih.gov/37045921/)]
42. Weatherbed J. OpenAI's regulatory troubles are only just beginning. *The Verge*. 2023. URL: <https://www.theverge.com/2023/5/5/23709833/openai-chatgpt-gdpr-ai-regulation-europe-eu-italy> [accessed 2024-07-20]

Abbreviations

AI: artificial intelligence
DM: diabetes mellitus
FKGL: Flesch-Kincaid Grade Level
FKRE: Flesch Reading Ease Score
GLM: generative language model
HTN: hypertension
LHL: low health literacy

Edited by K El Emam, B Malin; submitted 08.11.23; peer-reviewed by I Feinberg, M Lin; comments to author 07.04.24; revised version received 28.04.24; accepted 29.06.24; published 13.08.24.

Please cite as:

Spina A, Andalib S, Flores D, Vermani R, Halaseh FF, Nelson AM

Evaluation of Generative Language Models in Personalizing Medical Information: Instrument Validation Study

JMIR AI 2024;3:e54371

URL: <https://ai.jmir.org/2024/1/e54371>

doi: [10.2196/54371](https://doi.org/10.2196/54371)

PMID:

©Aidin Spina, Saman Andalib, Daniel Flores, Rishi Vermani, Faris F Halaseh, Ariana M Nelson. Originally published in JMIR AI (<https://ai.jmir.org>), 13.08.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR AI, is properly cited. The complete bibliographic information, a link to the original publication on <https://www.ai.jmir.org/>, as well as this copyright and license information must be included.

Original Paper

Perceptions of Family Physicians About Applying AI in Primary Health Care: Case Study From a Premier Health Care Organization

Muhammad Atif Waheed¹, MBBS, MRCS, MRCGP, DPD, MBA; Lu Liu², PhD

¹Primary Health Care Corporation, Doha, Qatar

²Bath Business School, Bath Spa University, Bath, United Kingdom

Corresponding Author:

Muhammad Atif Waheed, MBBS, MRCS, MRCGP, DPD, MBA

Primary Health Care Corporation

Al Minna Street (B Ring Road)

Doha, 26555

Qatar

Phone: 974 33015895

Email: dratifwaheed@gmail.com

Abstract

Background: The COVID-19 pandemic has led to the rapid proliferation of artificial intelligence (AI), which was not previously anticipated; this is an unforeseen development. The use of AI in health care settings is increasing, as it proves to be a promising tool for transforming health care systems, improving operational and business processes, and efficiently simplifying health care tasks for family physicians and health care administrators. Therefore, it is necessary to assess the perspective of family physicians on AI and its impact on their job roles.

Objective: This study aims to determine the impact of AI on the management and practices of Qatar's Primary Health Care Corporation (PHCC) in improving health care tasks and service delivery. Furthermore, it seeks to evaluate the impact of AI on family physicians' job roles, including associated risks and ethical ramifications from their perspective.

Methods: We conducted a cross-sectional survey and sent a web-based questionnaire survey link to 724 practicing family physicians at the PHCC. In total, we received 102 eligible responses.

Results: Of the 102 respondents, 72 (70.6%) were men and 94 (92.2%) were aged between 35 and 54 years. In addition, 58 (56.9%) of the 102 respondents were consultants. The overall awareness of AI was 80 (78.4%) out of 102, with no difference between gender ($P=.06$) and age groups ($P=.12$). AI is perceived to play a positive role in improving health care practices at PHCC ($P<.001$), managing health care tasks ($P<.001$), and positively impacting health care service delivery ($P<.001$). Family physicians also perceived that their clinical, administrative, and opportunistic health care management roles were positively influenced by AI ($P<.001$). Furthermore, perceptions of family physicians indicate that AI improves operational and human resource management ($P<.001$), does not undermine patient-physician relationships ($P<.001$), and is not considered superior to human physicians in the clinical judgment process ($P<.001$). However, its inclusion is believed to decrease patient satisfaction ($P<.001$). AI decision-making and accountability were recognized as ethical risks, along with data protection and confidentiality. The optimism regarding using AI for future medical decisions was low among family physicians.

Conclusions: This study indicated a positive perception among family physicians regarding AI integration into primary care settings. AI demonstrates significant potential for enhancing health care task management and overall service delivery at the PHCC. It augments family physicians' roles without replacing them and proves beneficial for operational efficiency, human resource management, and public health during pandemics. While the implementation of AI is anticipated to bring benefits, the careful consideration of ethical, privacy, confidentiality, and patient-centric concerns is essential. These insights provide valuable guidance for the strategic integration of AI into health care systems, with a focus on maintaining high-quality patient care and addressing the multifaceted challenges that arise during this transformative process.

(JMIR AI 2024;3:e40781) doi:[10.2196/40781](https://doi.org/10.2196/40781)

KEYWORDS

AI; artificial intelligence; perception; attitude; opinion; surveys and questionnaires; family physician; primary care; health care service provider; health care professional; ethical; AI decision-making; AI challenges

Introduction

Background

There is no universal definition for artificial intelligence (AI) [1]. AI has been defined in the literature as the branch of applied computer sciences in which algorithms are designed and are intended to perform different tasks while mimicking human intelligence [2]. It has been further defined as technologies that not only mimic human intelligence but can surpass them [3].

The global AI market in health care is projected to reach US \$27.6 billion by 2025 [4]. One of the studies estimated that AI could save US \$150 billion per annum by 2026 [5]. The human resource crisis in health care is already on the rise [6]. The global shortage of health care workers is approximately 17.4 million. Approximately, 50% of existing jobs will be in jeopardy or obsolete in 20 years [7].

Primary care is where AI would be most used in terms of opportunities on the broadest scale, where its power and future would be realized [8]. Moreover, AI has been regarded as a transformational force in the health care sector due to its impact on key stakeholders, such as primary care physicians, patients, systems, and financiers. AI will significantly impact many dimensions of clinical practice in the coming years, as machine learning (ML) and deep learning (DL) continue to hasten and will bring many advantages to both patients and clinicians [9]. There is not much literature on the impact of AI on employees, as little effort has been made empirically to study its impact [10]. Moreover, it is essential to know if AI is beneficial or detrimental to employees, as assisted AI, augmented AI, and autonomous AI have different implications on employee's roles. Physicians are required to adjust their roles accordingly as AI is modeling practices nowadays, and if physicians fail to adjust their roles, it can lead to detrimental effects on overall patient care [11]. Physicians must be prepared to embrace the changes that AI will bring to their roles and to lead this change themselves. Furthermore, primary care physicians in health care are the main stakeholders and the most crucial, valuable, highly knowledgeable, and skilled human resource. Hence, it is essential to understand family physicians' overall perception of AI application in primary care to develop organizational policies, modernize information technology infrastructure, develop AI literacy among physicians, establish and modify data privacy, data confidentiality, and code of ethics for the successful adoption and implementation of technology to gain a competitive advantage.

Aims and Objectives

This study assessed the role of AI in the management and practices of the Primary Health Care Corporation (PHCC) in Qatar, emphasizing its fundamental potential in health care management. The primary objective was to evaluate the impact of AI in improving health care practice at the PHCC. In addition, the study sought to determine the role of AI in managing health

care tasks and assess its impact on family physicians' job roles. Moreover, this research also examined the challenges and ethical ramifications associated with introducing AI in primary care services at the PHCC.

Literature Review

Understanding AI

AI is related to developing machines that mimic human cognitive processes such as learning, reasoning, and self-correction [3,12] and to performing tasks similar to a human mind [1]. It involves applying theoretical principles and the operation of applicable operating models to automate intellectual behaviors [13]. AI includes new concepts and solutions to address complex challenges [14]. In the field of medicine, AI introduces novel concepts such as a *digital physician*, reshaping the landscape [15].

Conceptualization of AI

The core of the AI system comprises neural-like elements, which are interconnected growing networks similar to a human brain that is active, associative, and homogenous with the ability to perceive, apprehend, and save information, enabling the system to learn, train, reason, and classify data to locate various patterns and connections to control external modalities [16]. To understand AI, it is essential to understand 2 key forms of AI: general and narrow AI. General AI refers to a machine's ability to perform any intellectual task performed by a human, whereas narrow AI algorithms are designed for a limited task. Health applications using AI are generally of the narrow type. AI subfields in the health care sector are expert systems, automation of robotic processes, natural language processing, ML, and DL [6]. An example of an expert system is clinical decision support systems. The growth of AI in health care has been possible in recent decades due to the faster computer processing of data and data collection. Large amounts of data collection have been possible due to widespread electronic health records, mobile health, telehealth, and the Internet of Things. Improvements in natural language processing, ML, and DL have made AI possible up to the stage where it mimics human intelligence, fueling active discussion in the literature on whether AI can replace human doctors in the future [17].

AI and Health Care

The rapid growth of technology in health care has become a catalyst for evidence-based practice, and the integration of AI holds significant potential for improving health care service delivery [18]. The perceptions of AI's impact, coupled with a deep understanding of the knowledge and interests of family physicians within primary care settings, is pivotal for the successful implementation of AI-based applications. Surprisingly, in an extensive survey across 4 of Saudi Arabia's largest hospitals, a general lack of AI knowledge was evident among 250 doctors, nurses, and technicians [19]. This signifies the importance of addressing knowledge gaps to harness the benefits of AI effectively.

A gender-specific perspective on AI knowledge was highlighted in a study among 387 medical students in India, revealing that, although female students exhibited little initial AI knowledge, they displayed heightened interest in the field [20]. Moreover, AI was perceived to play a major role in health care service delivery in the future. This gender dimension adds nuance to the broader understanding of AI adoption in health care.

The projected role of AI in future health care service delivery is significant. Examples abound, such as AI therapy, which is a web-based course developed by the University of Sydney that uses cognitive behavioral therapy to help patients with social anxiety disorder [21]. In pathology, a DL-based convolution neural network achieved performance comparable with that of a human pathologist for detecting metastatic breast cancer in tissue slides from a lymph node biopsy. Similarly, the convolutional neural network-based system was more precise and accurate for a tissue slide-based scoring system to predict a decline in kidney function than a traditional pathologist [22]. In ophthalmology, the Food and Drug Administration-approved DL system has been used to detect diabetic retinopathy. Similarly, the DL system has been developed and evaluated to diagnose and classify cataracts in pediatric patients based on slit-lamp examination images, glaucoma based on retinal nerve fiber layer or visual field, and keratoconus based on Scheimpflug tonometry [22]. Physical robots are becoming more sophisticated as AI is incorporated into their operating systems and are likely to show the same intelligence level as other AI applications [23]. Moreover, surgical robots are also used in minimally invasive procedures such as in urological; gynecological; and ear, nose, and throat surgeries. In recent advances in AI applications, IBM's Watson is an aiding tool for physicians to detect cardiovascular diseases and cancer [21]. The IBM Watson system can search and analyze data from a wide range of sources, surpassing human physicians' capacity in knowledge [22]. Similarly, the picture archiving and communication system can detect signs of diseases from chest x-ray, ultrasound, magnetic resonance imaging, and computed tomography (CT) scan by contextualizing data from past images, clinical reports, and laboratory studies [24]. Opportunistic health care management is the provision of health services or interventions that are not planned but rather is an opportunity to address a health care need or an issue such as smoking cessation, screening for hypertension, prediabetes, and diabetes during a routine medical consultation. In the United Kingdom, for example, as a public health policy, "Making Every Contact Count" requires health care professionals to provide such interventions [25]. This role of the physician can be assisted by AI. AI can be used for opportunistic screening for diabetic retinopathy [26] and opportunistic screening of low bone density using contrast and noncontrast CT examinations [27]. AI algorithms identify minor or subclinical electrocardiogram abnormalities linked to a higher risk of developing left ventricular systolic dysfunction in the future [28]. During the COVID-19 pandemic, AI performed exceptionally well in the diagnosis, prognostic evaluation, epidemic forecasting, and drug discovery processes [29].

Building upon this extensive literature review, the following hypotheses were formulated:

- Hypothesis 1: perceived knowledge of AI among family physicians within the PHCC does not vary significantly based on age and gender.
- Hypothesis 2: family physicians at the PHCC perceive AI to have a positive impact on enhancing health care practices.
- Hypothesis 3: family physicians at the PHCC perceive AI to positively influence their roles in opportunistic health care management.

AI in Primary Care

The rapid advancement of AI technology has brought about transformative changes in health care management and has the potential to revolutionize various aspects of medical practice.

Kueper et al [30] conducted a scoping review of the literature on AI's application in primary care, highlighting the evolving landscape of AI adoption. This study showed a shift from traditional expert systems to more sophisticated approaches, particularly supervised ML, mirroring the rapid advances in AI technologies. This paradigm shift holds profound significance for health care management, as AI gains increasing recognition as an asset in supporting health care professionals to make well-informed clinical decisions, especially when managing chronic conditions in high-income countries.

AI can assist with various health care tasks in primary care settings. AI predictive analytics tools have proven their efficacy in managing health care tasks [31]. These tasks include maintaining precise medical records, scheduling, inventory management, cost tracking, health promotion, clinical diagnosis, treatment planning, and developing care management plans. AI has the potential to enhance health care outcomes by streamlining operations within the health care system. Current academic literature presents an increasing trend in AI use in primary care and its positive effect on health care tasks, particularly in clinical management responsibilities [18]. The emergence of AI in electronic health record systems, which are extensively used today, has proven to be highly effective. AI-based clinical decision support systems have continued to evolve. Clinical decision support systems assist physicians and enhance patient safety by preventing dosage errors, drug duplication and presenting information on drug and drug interactions. Moreover, these systems help physicians to adhere to clinical guidelines, order and interpret laboratory results, issue prompts and alerts for abnormal results, suggest follow-up actions, render treatment reminders and provide support for clinical and diagnostic coding [32]. Early diagnosis and treatment are essential for improving health outcomes. AI has shown its effectiveness in assisting doctors with image-based diagnoses of skin conditions and in proactively identifying patients at risk of developing dementia [15]. In England's National Health Services, innovative applications of AI range from triage and symptom assessment to the automatic coding of clinical data, both in primary and community health care settings, thereby supporting personalized care management. This integration not only improves the quality of care but also saves valuable time for physicians, allowing them to focus more on providing personalized patient care. While the transformative potential of AI in clinical roles is evident, its impact on administrative functions has been relatively less explored [23].

Nevertheless, AI can play a vital role in optimizing administrative processes such as insurance collection, clinical reporting, medical billing, sales cycle management, and medical record management, ultimately contributing to more efficient health care operations and resource allocation. AI systems can perform routine operational tasks such as maintenance system management, accounting, and information inquiry much better and faster than human workers. AI-enabled chatbots and nursing robots can significantly improve operational process efficiency and reduce medical cost [33].

Collective evidence from the literature strongly suggests that AI holds promise as a constructive tool for managing various health care tasks at the PHCC. Its integration is expected to lead to improved clinical decision-making, operational efficiency, and ultimately contribute to enhanced patient care.

Accordingly, the following hypotheses were developed:

- Hypothesis 4: AI is perceived to play a constructive role in managing various health care tasks in PHCC.
- Hypothesis 5: family physicians at the PHCC perceive AI as having a positive impact on their clinical management responsibilities.
- Hypothesis 6: family physicians at the PHCC perceive AI as having a positive impact on their administrative management tasks.
- Hypothesis 7: AI is perceived to significantly improve the operational processes at the PHCC.

AI and Physicians

Whether AI will eventually replace physicians or complement them is still being debated, but it will significantly impact health care management activities and service delivery. The study by Ahuja [22] investigated whether AI will augment the physician's role or eventually replace them using a quantitative survey methodology. The key finding was that AI would eventually replace radiologists in the field of radiology. This is because AI is more efficient and can handle and interpret millions of images in seconds. AI could interpret CT scans during the COVID-19 pandemic with 96% accuracy in just 20 seconds [34]. However, it has limitations, such as the inability to engage in complex interactions (ie, communication) with patients, failing to reassure patients, and to convey empathy. The study by Sarwar et al [35] concluded a positive attitude toward AI by taking the opinions of 487 pathologists from 54 countries regarding AI. However, the majority also had concerns regarding AI replacing their jobs. AI can triage cases as benign or malignant cases to pathologists, increasing diagnostic efficiency and accuracy and automated reporting, freeing 40% of pathologist time by reducing the workflow [36]. Esteva et al [37] conducted a comparative study testing 21 board-certified dermatologists against a convoluted neural network-trained system fed with 129,450 clinical images to the system. This study concluded that the CNN outperformed dermatologists in terms of both sensitivity and specificity. The study by Karches [38] argued that human physician judgment would remain better than that of AI in a primary care setting, as AI cannot adjust to recommendations according to individual patients' needs. However, it cannot fine-tune its perception based on the patient's history and examination, which appears to be a human-only

ability. The study by Amisha et al [21] contends that machines cannot gather cues that only a physician can do during a patient-physician encounter. The machine cannot translate human traits, such as empathy, creativity, imagination, critical thinking, emotional intelligence, and interpersonal communication, both analytically and logically. According to the study by Meskó et al [39], the human physician is inevitable as empathy, communication, and human touch are included in the entire treatment process, which AI cannot provide; hence, AI will only be a helpful cognitive assistant. Physicians and nurses provide care to patients in an empathetic and compassionate environment that robotic physicians and nurses will not be able to do, as they lack the human characteristics of compassion [40]. Trust, empathy, and compassion are widely acknowledged as the core principles of effective health care [41]. Empathetic care enhances patient satisfaction.

Accordingly, the following hypotheses were developed.

- Hypothesis 8: the application of AI in health care tasks is perceived to lead to improved health care service delivery at the PHCC.
- Hypothesis 9: family physicians at the PHCC believe that AI is less likely to replace their current job roles.
- Hypothesis 10: family physicians at the PHCC perceive that AI decision-making does not surpass the judgment process of human physicians.
- Hypothesis 11: family physicians believe that the introduction of AI at the PHCC reduces patient satisfaction.

AI and Human Resource Management

AI is promising for closing the care gap in resource-poor settings, as digital health is widening. AI can address human resource shortages by ameliorating diagnostics, administrations, big analytics, and health care decisions [39]. AI and big data have significant impacts on strategic human resource management. Its digital transformation improves business processes [42] through the employee recruitment process (hiring and selection) and performance evaluation by providing real-time and accurate data and positively impacting staff retention [43]. AI is reported to reduce costs by providing evidence-based and affordable care to patients [39]. Furthermore, this will improve the overall quality of care. The actual economic impact of AI on health care is undetermined due to the methodological deficiencies in the literature analyzed in a systematic review [44]. Human resource management confers a competitive advantage through employment management by placing capable and highly committed workers and incorporating structural, cultural, and personal techniques [45]. Human resource management aims to manage human capital in modern organizations, which is the most vital asset. Instead of focusing on power, human resource management should invest more in employee training and development because it is a significant source of innovation and development [46]. Mutual complementation of humans and machines creates more value for the organization, as machines help data interpretation and analysis, and humans in innovation and social interaction. AI frees employees from repetitive tasks, but at the same time, it also needs the development of higher collaborative competencies among employees. Firms in the future would need

new human resource plans with the need to develop policies by reviewing the need for structural changes and capacity models, and recent reforms would be required for enterprise human resource management. Furthermore, a negative attitude toward AI needs to be addressed among employees by properly engaging them and studying AI in depth [46].

Accordingly, the following hypothesis was developed:

- Hypothesis 12: the introduction of AI is perceived to assist and enhance human resource management practices at the PHCC.

AI Challenges, Risks, and Ethical Ramifications

The study by Lai et al [1] argued that there are many concerns regarding AI. These include the fuzzy notion of AI, health data confidentiality issues, growth in AI knowledge, international competition, and disruption of the patient-physician relationship. Furthermore, the diagnosis and decision-making landscape is expected to change for both physicians and patients, and these developments would impact the entire health care system. The implementation of AI will be another challenge, and AI must add value and should support and not subvert the patient-physician relationship, as health care is a social endeavor based on human interactions. If AI is implemented correctly, emotional and cognitive spaces would open for physicians; however, if implemented incorrectly, it will have severe consequences [8]. The existing information technology infrastructure might be outdated to adopt AI systems, which will require careful review before implementation. The adoption of AI-based technologies by resource constrained countries can be wider, as they will be more open to policy changes compared to resource-rich countries [39]. With the introduction of AI technology, the patient-physician relationship will change significantly. The hierarchy will still be in place and the patient-physician relationship will be more just than ever before but patients' autonomy is still a question. Similarly, the development of standards for collecting data and testing, which stakeholders, clinicians, industry, and scientists should lead, will be challenging.

Accordingly, the following hypotheses were developed:

- Hypothesis 13: the implementation of AI is not expected to undermine the patient-physician relationship from the family physician's perspective at the PHCC.
- Hypothesis 14: family physicians' perceptions of challenges and ethical ramifications when introducing AI at the PHCC do not significantly differ based on age and gender.

Methods

Overview

This study adopted a cross-sectional research design, using a quantitative approach. The primary focus of a quantitative methodology is to identify the relationship between variables and to accept or reject connections or linkages between these variables [47]. Moreover, it reduces bias probabilities, as the researcher is independent of the respondents, both physically and emotionally, and establishes standardization of investigation and interpretation rather than situational analysis.

Participants and Data Collection

The PHCC in Qatar is the main provider of primary health care services via its 28 health centers, scattered across all regions of Qatar. There are 724 family physicians working in the organization. On the basis of a CI of 95%, an expected proportion of 0.5, and a margin of error of 5, the sample size was 252. The questionnaire was sent via PHCC intranet email by the operations department of the PHCC to all family physicians for 4 weeks in March 2021, with reminder emails sent after the second and third week. A total of 132 physicians participated in the study. Among them, 102 questionnaires were fully completed that were eligible for data analysis. Incomplete questionnaires were not included because all questions were eligible for the hypothesis test. The response rate was 14.1% (102/724).

The demographic characteristics of participants who were early and late responders were analyzed by splitting the data into 2 groups based on the date of response. The analysis showed that early responders were similar to late responders in age ($P=.15$), gender ($P=.99$), working status ($P=.33$), and licensed years ($P=.11$). Although the response rate was low, it can be assumed that the nonresponse bias was minimal.

Data Analysis

Data collected from participants using Survey Monkey software (Symphony Technology Group) were transferred to SPSS (version 27; IBM Corp) for statistical analysis. The information was codified using the data statistics editor in SPSS. The Likert scale had 5 categories: strongly agree, agree, neutral, disagree, and strongly disagree. The Likert scale data were forward scored on a numeric scale of 1 to 5 to facilitate statistical analysis.

Descriptive and inferential statistical models were used to analyze the survey results. Nonparametric tests were chosen because the data were not normally distributed, as confirmed using the Kolmogorov test. The Shapiro-Wilk test is commonly used for sample sizes <50 , while the Kolmogorov approach is used for sample sizes >50 to assess the normality of the data distribution [48].

The Spearman rho correlation coefficient, which assesses the 2-way linear relationship between 2 variables, was used to determine whether the application of AI in health care tasks is perceived to lead to improved health care service delivery at PHCC. The Spearman rho correlation value ranges from +1 to -1, where 0 represents no relationship, +1 indicates a perfect positive correlation, and -1 indicates a perfect negative correlation [49].

The chi-square goodness-of-fit test, comparing expected and observed values in categorical variables [50], was used to assess family physicians' perceptions of AI's role in job replacement, human resource management, and patient-physician relationships. The chi-square test of homogeneity, comparing proportions between ≥ 2 groups [51], was used to examine variations in AI knowledge and challenges, as well as ethical ramifications by age and gender. To compare column proportions by age and gender groups, multiple corrections were made using the Bonferroni correction.

The 1-sample Wilcoxon test was used to assess perceptions of AI's positive impact on health care practices, clinical and administrative tasks, and patient satisfaction at the PHCC. This test, an alternative to the standard 1-sample t test, is assumed to be more sensitive to the sign test, measuring positive and negative ranks for testing significance using the hypothesized median set as neutral (0) when testing these hypotheses [52].

Validation

The survey questionnaire, comprising 47 questions, underwent a systematic process of piloting, testing, and validation to eliminate potential ambiguity for the respondents. Primarily using the Likert scale, it covered five main constructs: (1) demographics (4 items), exclusively designed to capture participant data without internal consistency measurement; (2) family physician's knowledge and perspective on clinical management of AI (11 items, Cronbach $\alpha=0.873$); (3) family physician perspective on administrative management of AI (10 items, Cronbach $\alpha=0.916$); (4) family physician's perspective on public health management of AI (9 items, Cronbach $\alpha=0.930$); and (5) family physicians' perspectives on AI challenges, ethical ramifications, and impact on job roles (13 items, Cronbach $\alpha=0.744$). The overall Cronbach α score for the research instrument, excluding demographics, was 0.937, indicating exceptional reliability per the established standard [53]. The respondents took mean time of 12 (SD 9) minutes to complete the survey.

Ethical Considerations

This paper was developed from the first author's (MAW's) dissertation that he completed with the University of Liverpool in partial fulfillment of the requirements for a master's degree when the second author (LL) was the supervisor. The original research project was approved by both the University of Liverpool, United Kingdom Research Ethics Committee, and PHCC, Qatar Research Subcommittee (approval no: PHCC/DCR/2020/07/079). Permission was obtained by sending a questionnaire via the intranet email from the organization. In the web-based survey questionnaire, an initial page was provided to participants to explain the nature, purposes, and expected duration of the research. Moreover, it was ensured to the participants that this study was entirely voluntary, their data would be dealt with in the strictest confidential manner, and no information would be collected to identify them.

Results

Descriptive Statistics

Descriptive statistics were used to summarize the research findings using the frequency and percentage of responses. The responses on the Likert scale were collapsed and recategorized into 3 main groups: agree, neutral, and disagreed.

Demographic Data

The demographic data are summarized in [Table 1](#). The majority of respondents (72/102, 70.6%) were men, 94 (92.2%) out of 102 were in the age group of 35 to 54 years, and 58 (56.9%) out of 102 worked as consultants.

Table 1. Participants' demographic data (N=102).

Characteristics	Values, n (%)
Age groups (years)	
25-34	1 (1)
35-44	44 (43.1)
45-54	49 (48)
55-64	8 (7.8)
Gender	
Men	72 (70.6)
Women	30 (29.4)
Working status	
General practitioner	27 (26.5)
Pediatrician	1 (1)
Associate specialist	7 (6.9)
Consultant	58 (56.9)
Senior consultant	7 (6.9)
Manager	1 (1)
Executive	1 (1)
Licensed years	
1-2	2 (2)
2-5	16 (15.7)
5-10	29 (28.4)
10-20	31 (30.4)
20-30	19 (18.6)
30-40	5 (4.90)

Perceived Knowledge of AI

Of the 102 family physicians surveyed for AI awareness, 7 (6.9%) out of 102 were extremely aware, 18 (17.6%) out of 102 were very aware, 55 (53.9%) out of 102 were somewhat aware,

20 (19.6%) out of 102 were not so aware, and 2 (2%) out of 102 had no awareness. Overall, AI awareness among PHCC physicians was 78.4% (80/102). The results are summarized in [Table 2](#).

Table 2. Perceived knowledge of artificial intelligence (AI; N=102).

Perceived knowledge of AI	Values n (%)
Extremely aware	7 (6.9)
Very aware	18 (17.6)
Somewhat aware	55 (53.9)
Not so aware	20 (19.6)
Not at all aware	2 (2)
Overall awareness	80 (78.4)

Family Physicians' Perspective on Clinical and Administrative Role of AI in Health Care Management

[Table 3](#) depicts the perspective of family physicians on the clinical and administrative role of AI in health care management. Most of the respondents (73/102, 71.6%) acknowledge the potential of AI in triage, while 60 (58.8%) out of 102 believe in its efficacy for assisting in emergency case management.

Regarding clinical assessment and diagnostic management tasks, 69.6% (71/102) agree with the assistive role of AI, with 55 (53.9%) out of 102 of physicians foreseeing its capability to surpass conventional methods of diagnostic report management. Furthermore, 81 (79.4%) and 56 (54.9%) out of 102 of physicians believe in AI's assistance in medication management requirements and improving patient treatment compliance, respectively. On the administrative role of AI, most physicians

(86/102, 84.3%) perceive AI as managing health care performance by enhancing information dissemination, while 78 (76.5%) out of 102 anticipate improved efficiency in health care administrative activities. The positive perceptions extend to the care management systems, with 83 (81.4%) out of 102

agreeing on AI's improving them and 79 (77.5%) out of 102 endorsing its ability to reduce medical errors. This collective optimism highlights the potential transformative impact of AI in enhancing the clinical and administrative roles of family physicians and, hence, health care delivery.

Table 3. Family physicians' perspective on the clinical and administrative management role of artificial intelligence (AI; N=102).

	Agree, n (%)	Neutral, n (%)	Disagree, n (%)
AI on clinical management			
AI can assist in triage	73 (71.6)	23 (22.5)	6 (5.9)
AI can assist in managing emergency cases	60 (58.8)	27 (26.5)	15 (14.7)
AI will assist in clinical assessment and diagnosis management tasks easy	71 (69.6)	24 (23.5)	7 (6.9)
AI will supersede the conventional methods of diagnostic reports management	55 (53.9)	30 (29.4)	17 (16.7)
AI will improve clinical judgment process	65 (63.7)	28 (27.5)	9 (8.8)
AI has the potential for the task management and clinical investigation and data storage	86 (84.3)	15 (14.7)	1 (1)
AI will assist to follow clinical pathways	82 (80.4)	19 (18.6)	1 (1)
AI will assist in medication management requirements	81 (79.4)	19 (18.6)	2 (2)
AI integration will make patients treatment compliance better	56 (54.9)	35 (34.3)	11 (10.8)
AI enhances overall management of patient care	71 (69.6)	29 (28.4)	2 (2)
AI on administrative management			
AI integration will help in improving care management systems	83 (81.4)	18 (17.6)	1 (1)
AI can make effective plans to reduce medical errors	79 (77.5)	22 (21.6)	1 (1)
AI will help in managing health care performance by improving information dissemination	86 (84.3)	15 (14.7)	1 (1)
AI will be more helpful in management of service provision	78 (76.5)	22 (21.6)	2 (2)
AI integration will make health care administrative activities more robust and successful	78 (76.5)	20 (19.6)	4 (3.9)
AI helps in financial planning and management	75 (73.5)	24 (23.5)	3 (2.9)
AI assists in health care policy making	67 (65.7)	24 (23.5)	11 (10.8)
AI has potential for planning treatment care pathways	73 (71.6)	23 (22.5)	6 (5.9)
AI will assist human resource management (recruitment and retention)	65 (63.7)	29 (28.4)	8 (7.8)
AI introduction will be advantageous to administrative staff	74 (72.5)	23 (22.5)	5 (4.9)

Family Physicians' Perspective on the Role of AI in Public Health Management

Table 4 presents the family physicians' perspectives on the role of AI in public health management. Family physicians overwhelmingly supported the integration of AI in public health, with 80 (78.4%) out of 102 respondents endorsing its role in organizing tasks for public health awareness and 84 (82.4%) out of 102 endorsing its role in managing public health surveillance. Interestingly, 85 (83.3%) out of 102 agreed on

AI's efficacy in providing disease reports for disease prediction and management. A significant majority (86/102, 84.3%) perceived AI as a valuable tool for opportunistic health care screening. Moreover, 78 (76.5%) out of 102 believed in AI's effectiveness during epidemics and 72 (70.6%) out of 102 agreed that it aids in managing health care logistics and reducing costs during pandemics. These findings highlight the positive perception of AI's multifaceted benefits in enhancing public health strategies and outcomes.

Table 4. Family physicians' perspectives on the role of artificial intelligence (AI) in public health management (N=102).

	Agree, n (%)	Neutral, n (%)	Disagree, n (%)
AI on public health management			
AI is beneficial for organizing tasks for public health awareness	80 (78.4)	21 (20.6)	1 (1)
AI helps in disease screening and monitoring	84 (82.4)	17 (16.7)	1 (1)
AI is an efficient tool for assessing and managing risks to public health	80 (78.4)	20 (19.6)	2 (2)
AI has the potential for providing reports for disease prediction and disease management	85 (83.3)	17 (16.7)	0 (0)
AI may be considered by physicians as a beneficial tool in managing public health surveillance	84 (82.4)	18 (17.6)	0 (0)
AI introduction in health care management will make opportunistic health care screening easier	86 (84.3)	16 (15.7)	0 (0)
AI is effective tool in managing quality of care in epidemics	78 (76.5)	21 (20.6)	3 (2.9)
AI is an efficient tool in disease containment projects planning	73 (71.6)	27 (26.5)	2 (2)
AI will help in managing health care logistics and reduce cost during pandemics	72 (70.6)	30 (29.4)	0 (0)

Family Physicians' Perspective on AI Challenges and Ethical Ramifications in Health Care and Impact on Their Job Roles

Table 5 shows that family physicians expressed concerns about AI challenges, ethical ramifications in health care, and their impact on their job roles. A majority (61/102, 59.8%) worried about patient confidentiality due to potential hacking of AI-managed health care records; similarly, 61 (59.8%) out of 102 were concerned about the risk to organizations' confidential data. Regarding decision-making, 69 (67.6%) out of 102 acknowledged potential conflicts with humans due to differences in decision-making and 80 (78.4%) out of 102 expressed concern about AI lacking emotional input. Patient satisfaction was a concern for 76 (74.5%) out of 102 due to the absence of emotions in AI-driven decisions. In addition, 65 (63.7%) out of 102 believed AI's clinical judgment may be inferior to that

of physicians. While 42 (41.2%) out of 102 agreed AI could be accountable in malpractice cases, 89 (87.3%) out of 102 emphasized the need for AI training for health care managers and staff. However, 33 (32.4%) out of 102 found learning AI challenging for health care staff. Family physicians expressed nuanced views on AI's impact on their roles. The majority (74/102, 72.5%) believed that AI cannot replace their jobs, with 53 (52%) out of 102 asserting that it will not undermine the patient-physician relationship. A total of 35 (34.3%) out of 102 were open to using AI in medical decisions in the future. These findings demonstrated family physicians' perceived AI risks, such as data privacy, confidentiality, the decision-making process of AI, its accountability in cases of malpractice, and the need for training to learn AI. Moreover, it also highlighted a balanced perspective on AI's role, emphasizing AI augmenting the roles of family physicians rather than replacing them.

Table 5. Family physicians' perception on artificial intelligence (AI) challenges and ethical ramifications and impact on their job role (N=102).

	Agree, n (%)	Neutral, n (%)	Disagree, n (%)
AI challenges and ethical ramifications			
Management of health care records through AI may threaten patient confidentiality due to hacking	61 (59.8)	32 (31.4)	9 (8.8)
Management through AI may threaten health care organizations confidential data due to hacking	61 (59.8)	30 (29.4)	11 (10.8)
Management of health care operations involving AI may conflict with humans due to difference in decision-making	69 (67.6)	26 (25.5)	7 (6.9)
Decision-making process by AI in health care encounters lacks emotional input	80 (78.4)	16 (15.7)	6 (5.9)
Management of decision-making process through AI may decrease patient satisfaction due to lack of emotions	76 (74.5)	18 (17.6)	8 (7.8)
Patients' satisfaction is decreased with inclusion of AI in decision-making process management	47 (46.1)	42 (41.2)	13 (12.7)
Process of clinical judgment by AI might be inferior to that made by physicians	65 (63.7)	26 (25.5)	11 (10.8)
In case of malpractice AI integration in decision-making process can be held accountable	42 (41.2)	39 (38.2)	21 (20.6)
Health care managers and staff will require training in AI-based operations	89 (87.3)	12 (11.8)	1 (1)
Management of health care processes through AI are hard to learn for health care staff	33 (32.4)	40 (39.2)	29 (28.4)
AI could not replace physician job	74 (72.5)	17 (16.7)	11 (10.8)
AI would not undermine patient-physician relationship	53 (52.0)	28 (27.5)	21 (20.6)
AI will be used in making medical decision in future	35 (34.3)	43 (43.1)	23 (22.5)

Hypothesis

Table 6 illustrates a summary of the hypotheses tested, the statistical tests used, corresponding *P* values and key findings with their relevant implications. The perceived knowledge of AI among different age and gender groups (hypothesis 1) examined by using the chi-square test of homogeneity showed no statistical significance for the perceived knowledge of AI among family physicians within the PHCC based on age and gender groups. The awareness of the physicians who were men was (60/72, 83%), and that of the women awareness was (20/30, 67%; *P*=.06). Similarly, regarding the awareness of AI between physicians aged 18 to 54 years (72/94, 77%) and aged >55 years (8/8, 100%) with *P*=.12. Licensed years and working status also had no statistical significance with awareness of AI (*P*=.50 and *P*=.51, respectively). Chi-square tests of homogeneity showed no significant differences across age and gender groups regarding 10 item, AI challenges and ethical ramifications

(hypothesis 14; *P*>.05). A 1-sample Wilcoxon signed-rank test confirmed a perceived positive role of AI in health care practice, task management, and operational processes at PHCC (hypotheses 2, 4, and 7; *P*<.001). In addition, a Spearman rho test demonstrated a moderate to strong correlation between health care tasks and health care service delivery (hypothesis 8; Spearman rho=0.679, *P*<.001). The analyses using a 1-sample Wilcoxon signed-rank test further supported the positive impact of AI on family physician opportunistic health and clinical and administrative roles (hypotheses 3, 5, and 6; *P*<.001), while anticipating a reduction in patient satisfaction (hypothesis 11; *P*<.001). Importantly, the results indicated that AI is not expected to negatively impact the patient-physician relationship (hypothesis 13; *P*<.001) and will not replace human physicians (hypothesis 11; *P*<.001). These findings provide valuable insights into the strategic integration of AI into health care settings.

Table 6. Summary of hypothesis testing using specific statistical tests (chi-square test of homogeneity and goodness-of-fit, 1-sample Wilcoxon signed-rank test, and Spearman rho), corresponding *P* values, key findings and their implications.

Hypothesis	Statistical test	<i>P</i> value	Key findings and implications
Hypothesis 1: perceived knowledge of AI ^a among family physicians within the PHCC ^b does not significantly vary based on age and gender groups.	Chi-square test of homogeneity	.12 for age; .06 for gender groups	No significant difference in AI perceived knowledge across age and gender groups.
Hypothesis 2: family physicians at the PHCC perceive AI to have a positive impact on enhancing health care practices.	1-sample Wilcoxon signed-rank test	<.001	Strong evidence is supporting the perceived positive role of AI in health care practice.
Hypothesis 3: family physicians at the PHCC perceive AI to positively influence their roles in opportunistic health care management.	1-sample Wilcoxon signed-rank test	<.001	Affirms the perceived positive influence of AI on opportunistic health care management roles.
Hypothesis 4: AI is perceived to play a constructive role in managing various health care tasks at the PHCC.	1-sample Wilcoxon signed-rank test	<.001	Strong evidence suggesting AI's perceived beneficial impact on health care task management.
Hypothesis 5: family physicians at the PHCC perceive AI to have a positive impact on their clinical management responsibilities.	1-sample Wilcoxon signed-rank test	<.001	Indicates a perceived positive effect of AI on clinical management roles.
Hypothesis 6: family physicians at the PHCC perceive AI to have a positive impact on their administrative management tasks.	1-sample Wilcoxon signed-rank test	<.001	Provides evidence of AI's perceived positive influence on administrative roles.
Hypothesis 7: AI is perceived to significantly improve the operational processes at the PHCC.	1-sample Wilcoxon signed-rank test	<.001	Strong evidence supporting AI's perceived positive influence on health care operations.
Hypothesis 8: the application of AI in health care tasks is perceived to lead to improved health care service delivery at the PHCC.	Spearman rho test ^c	<.001	Moderate to strong positive correlation between perceived AI application in health care tasks and health care service delivery.
Hypothesis 9: family physicians at the PHCC believe that AI is less likely to replace their current job roles.	Chi-square goodness-of-fit test ^d	<.001	Strong evidence against the hypothesis of AI job replacement as perceived by family physicians.
Hypothesis 10: family physicians at the PHCC perceive that AI decision-making does not surpass the judgment process of human physicians.	One sample Wilcoxon signed-rank test	<.001	Strong evidence against the superiority of AI decision-making over human judgment as perceived by family physicians.
Hypothesis 11: the introduction of AI is believed to reduce patient satisfaction by family physicians at the PHCC.	One sample Wilcoxon signed-rank test	<.001	Strong evidence that AI has a negative impact on patient satisfaction as perceived by family physicians.
Hypothesis 12: the introduction of AI is perceived to assist and enhance human resource management practices at the PHCC.	Chi-square goodness-of-fit test ^e	<.001	Strong evidence supporting the idea that AI is perceived to assist in human resource management.
Hypothesis 13: the implementation of AI is not expected to undermine the patient-physician relationship from the family physician perspective of the PHCC.	Chi-square goodness-of-fit test ^f	<.001	Strong evidence against the hypothesis of AI is perceived to negatively impacting the patient-physician relationship.
Hypothesis 14: family physicians' perceptions of challenges and ethical ramifications when introducing AI at the PHCC do not significantly differ based on age and gender.	Chi-square test of homogeneity	>.05	No significant differences in perceived challenges and ethical ramifications among age and gender groups.

^aAI: artificial intelligence.

^bPHCC: Primary Health Care Corporation.

^cCorrelation coefficient of health care tasks and health care service delivery was Spearman rho=0.679 (moderate to strong correlation).

^d $\chi^2_2=71.1$; N=102.

^e $\chi^2_2=48.8$; N=102.

^f $\chi^2_2=16.6$; N=102.

Discussion

Principal Findings

The primary findings of this study offer valuable insights into the perceptions of PHCC family physicians in Qatar regarding

the integration of AI in the health care context. The overall awareness of AI among PHCC physicians in Qatar was 78.4% (80/102). Moreover, the proportion of physicians with very aware and extremely aware levels of AI was 24.5% (25/102), reflecting a robust understanding of AI technology. Critically, the statistical analysis did not reveal any meaningful variations

in perceived AI knowledge based on gender ($P=.06$) or age groups ($P=.12$). Similarly, the exploration showed no statistically significant correlations between AI awareness and factors such as years of licensure ($P=.50$) or current working status ($P=.51$). Similarly, no significant disparities in perceived AI challenges and ethical implications were identified among physicians of diverse age and gender groups ($P>.05$). Furthermore, the results highlight the affirmative role that physicians perceive AI might play in the enhancement of health care practices at the PHCC ($P<.001$), facilitating improved management of health care tasks ($P<.001$), optimizing operational processes ($P<.001$), and fostering effective human resource management ($P<.001$). Notably, AI was perceived to exert a beneficial influence on the multifaceted roles of family physicians in clinical ($P<.001$), administrative ($P<.001$), and opportunistic health care management ($P<.001$). It is crucial to highlight that the study findings indicate physicians' perception that AI decision-making does not supersede the clinical judgment process of human physicians ($P<.001$), and the introduction of AI is not anticipated to compromise the essential patient-physician relationship ($P<.001$). Moreover, from the perspective of family physicians, AI was less likely to displace their existing job roles ($P<.001$). However, the implementation of AI was expected to result in reduced patient satisfaction ($P<.001$).

Comparison With Prior Work

The overall awareness of AI among PHCC physicians stands at 78.4% (80/102), reflecting a significant level of perceived knowledge. This heightened awareness may facilitate the implementation of AI without substantial resistance [54]. This awareness level is notably higher than that in the study conducted by Oh et al [55], where only 5.9% of Korean medical students and doctors perceived a strong familiarity with AI, despite Korea's reputation as technologically advanced.

Consistent with the proposition found in the study by Lin et al [8], our findings indicate that PHCC physicians perceive AI as a transformative force in primary care. Importantly, our research affirms that from the physicians' perspective, AI is less likely to replace the role of the family physician and does not surpass the human physician decision-making process. This aligns with the literature, which asserts that AI enhances the diagnostic capability of family physicians rather than replacing their diagnostic intelligence [56].

Our study demonstrates that PHCC physicians perceive AI as a valuable tool for human resource management, positively impacting both employee retention and recruitment, which is consistent with the literature. Despite being a relatively novel concept, AI has the potential to streamline recruitment processes, leading to more efficient and high-quality employee selection [57]. Furthermore, AI's influence extends across key domains of human resource management, as indicated by its potential to enhance recruitment, placement, staff development, performance management, compensation management, human relations management, and strategic planning of human resources [58]. AI-based systems, such as those using automated recruitment tasks and reducing bias, hold promise for improving the efficiency and effectiveness of human resource functions.

The perception among PHCC physicians that AI improves operational processes and reduces the cost of care aligns with existing literature. Predictive analytics, including forecasting, enhance capacity management, resource use, and improvement in overall business processes, contributing to operational innovation in health care [59]. In addition, routine operational processes can be made quicker and more efficient through AI integration.

Although, nowadays, AI can demonstrate superior performance compared to physicians in certain specialties, such as dermatology (analysis of skin lesions), pathology (slide scanning), cardiology (electrocardiographic interpretation), and radiology (analysis of clinical images) [60], it is not perceived as surpassing the broader clinical decision-making process of human physicians. Patient satisfaction may be reduced due to AI's limitations in replicating human characteristics, such as empathy, compassion, and human touch [61], and complete acceptance of fully automated services remains a challenge. Nevertheless, AI's superiority in specialized domains underscores its potential to complement medical practitioners in specific areas.

This study highlights the perceived positive impact of AI on opportunistic health care management, which was evident particularly during the COVID-19 pandemic. The use of AI in tracking, prediction, contact tracing, early diagnosis, monitoring, and vaccine development highlights its crucial role in addressing pandemic health care challenges [62]. Approximately 36 countries have used AI- and ML-based applications for digital contact tracing to limit the spread of SARS-CoV-2 [63]. The Ministry of Public Health of Qatar has also adopted AI-based tools for contact tracing, and this exemplifies how AI can contribute to crisis management and safeguarding public health.

Given the perception of family physicians, this research establishes that AI integration positively affects PHCC service delivery, enhancing health care task management and care systems. AI will automate many administrative tasks where managers, administrators, and health care staff spend about 54% of their time on them [64]. Family physicians' clinical and administrative roles may benefit from AI integration, reducing administrative burdens and allowing them to focus on patient-centered care, increasing their professional fulfillment and reducing burnout [65]. AI's potential for disease prediction, digital health coaching, evidence-based clinical decisions, and medication management improvement holds promise for improving the quality of care provided.

PHCC physicians perceived ethical considerations surrounding AI, including informed consent, safety, transparency, biases, and data privacy, aligned with concerns found in the literature [66]. Notably, 41.2% (42/102) of participants in this study advocated AI's liability in cases of malpractice, reflecting the need for robust accountability mechanisms. Recent regulatory updates, such as the introduction of the Medical Device Regulation in Europe, reflect the evolving legal landscape of AI [66]. Policy makers should consider product liability, deterrence, and compensation as they navigate this dynamic terrain.

While the impact of AI on patient-physician relationships remains uncertain [67], our study concludes that from the physicians' perspective, AI will not subvert these relationships. However, careful and strategic planning is essential during AI implementation to prevent potential negative consequences. The balance between cost reduction, efficiency, and accuracy considerations while upholding patient-physician dynamics is of paramount importance.

Limitations and Further Research

This study produced compelling findings and will serve as a springboard for future researchers to replicate similar studies. However, it is critical to understand the limitations of a study because they reflect flaws that could influence the outcomes and conclusions [68]. First, it used a positivist paradigm that limits family physicians' richer perspectives in a broader context for applying AI in PHCC management and practices. Second, this study only included the PHCC, a single organization, and the response rate in this study was low despite sending 2 reminder emails to practicing family physicians at the PHCC. However, response rates have been declining in health care field-related surveys [69] and physicians' response rates have continued to decline [70].

This research can be replicated based on an interpretivist paradigm and by using semistructured interviews to obtain deeper insights and richer knowledge about the perception of family physicians regarding the application of AI in a primary care setting. Perhaps using mixed methods will provide a deeper understanding and add more rigor to research regarding the application of AI in primary care [71]. Future research can also examine the factors that lead to resistance to AI implementation in primary care. Moreover, it should include nurses' administrative, laboratory, pharmacy, and dental staff' perspectives on applying AI in primary care. Furthermore, the most crucial aspect is to have the patient perspective central to improvement in health care systems.

Conclusions

The findings from this study indicate that physicians hold a very positive perception regarding the integration of AI within

primary health care services at the PHCC, foreseeing potential enhancements in health care task management and overall service delivery. This perception extends to various dimensions of family physicians' job roles, encompassing clinical, administrative, and opportunistic health care management. The positive expectations regarding AI's impact also extend to operational processes, anticipating improved information dissemination, enhanced health care policy formulation, optimization of treatment care pathways, more effective human resource management, and strategic financial planning processes within the PHCC. During periods of epidemics and pandemics such as the COVID-19 pandemic, the public health management role of AI is well acknowledged by family physicians for disease screening, contact tracing, risk assessment, real-time monitoring, early diagnosis, vaccine development, and formulating efficient management strategies using AI's predictive and logistical prowess. It is important to note that AI is not perceived as a direct replacement for family physician roles, and its introduction is not anticipated to undermine the significant patient-physician relationship. Moreover, AI is not perceived as superior to the human judgment process. Although AI holds the potential to be a valuable augmentation tool for the roles of family physicians, as per their perspective, it enhances their efficiency and productivity. However, its implementation requires due diligence with a strategy that maintains the critical challenges associated with AI integration, such as concerns related to patient satisfaction, ethical considerations regarding AI accountability in cases of malpractice, and the utmost need to uphold data privacy and confidentiality, as highlighted in this study. The implementation of AI is expected to elevate care management systems, consequently enhancing the quality of care, while simultaneously streamlining costs. The perception-based insights from this study can guide future AI implementation strategies within the context of primary health care at the PHCC, helping to pave the way for a more informed and sustainable integration of this technology. This careful and patient-centered approach will be essential in unlocking the full potential of AI in improving health care delivery, while safeguarding the values and priorities that underpin the field of medicine.

Acknowledgments

I am grateful to LL, who is the coauthor and was the dissertation advisor, for her invaluable guidance and supervision throughout this research. I extend a special heartfelt thanks to Dr Lolwa Al Mannai for her unwavering support and motivation. I express my sincere gratitude to Dr Samya Ahmad Al Abdulla, the executive director of operations of Primary Health Care Corporation, for her encouragement, support, mentorship, and project approval, without which completing this project would not have been possible. Moreover, I extend my sincere appreciation to our colleague, Dr Hashim AlSayed Mohammed, and all the physicians who contributed to this project.

Data Availability

All data generated or analyzed during this study are included in this published paper.

Conflicts of Interest

None declared.

References

<https://ai.jmir.org/2024/1/e40781>

JMIR AI 2024 | vol. 3 | e40781 | p.585
(page number not for citation purposes)

1. Lai MC, Brian M, Mamzer MF. Perceptions of artificial intelligence in healthcare: findings from a qualitative survey study among actors in France. *J Transl Med* 2020 Jan 09;18(1):14 [FREE Full text] [doi: [10.1186/s12967-019-02204-y](https://doi.org/10.1186/s12967-019-02204-y)] [Medline: [31918710](https://pubmed.ncbi.nlm.nih.gov/31918710/)]
2. Filipović-Grčić L, Đerke F. Artificial intelligence in radiology. *Rad CASA Med Sci* 2019;537(46-47):55-59. [doi: [10.21857/y26kec3o79](https://doi.org/10.21857/y26kec3o79)]
3. Kar UK, Dash R. The future of health and healthcare in a world of artificial intelligence. *Arch Biomed Eng Biotechnol* 2018;1(1). [doi: [10.33552/abeb.2018.01.000503](https://doi.org/10.33552/abeb.2018.01.000503)]
4. Barbour AB, Frush JM, Gatta LA, McManigle WC, Keah NM, Bejarano-Pineda L, et al. Artificial intelligence in health care: insights from an educational forum. *J Med Educ Curric Dev* 2019;6:2382120519889348 [FREE Full text] [doi: [10.1177/2382120519889348](https://doi.org/10.1177/2382120519889348)] [Medline: [32064356](https://pubmed.ncbi.nlm.nih.gov/32064356/)]
5. Wahl B, Cossy-Gantner A, Germann S, Schwalbe NR. Artificial intelligence (AI) and global health: how can AI contribute to health in resource-poor settings? *BMJ Glob Health* 2018;3(4):e000798 [FREE Full text] [doi: [10.1136/bmjgh-2018-000798](https://doi.org/10.1136/bmjgh-2018-000798)] [Medline: [30233828](https://pubmed.ncbi.nlm.nih.gov/30233828/)]
6. Wiljer D, Hakim Z. Developing an artificial intelligence-enabled health care practice: rewiring health care professions for better care. *J Med Imaging Radiat Sci* 2019 Dec;50(4 Suppl 2):S8-14. [doi: [10.1016/j.jmir.2019.09.010](https://doi.org/10.1016/j.jmir.2019.09.010)] [Medline: [31791914](https://pubmed.ncbi.nlm.nih.gov/31791914/)]
7. Mesko B. Health IT and digital health: the future of health technology is diverse. *J Clin Transl Res* 2018 Dec 17;3(Suppl 3):431-434 [FREE Full text] [Medline: [30873492](https://pubmed.ncbi.nlm.nih.gov/30873492/)]
8. Lin SY, Mahoney MR, Sinsky CA. Ten ways artificial intelligence will transform primary care. *J Gen Intern Med* 2019 Aug;34(8):1626-1630 [FREE Full text] [doi: [10.1007/s11606-019-05035-1](https://doi.org/10.1007/s11606-019-05035-1)] [Medline: [31090027](https://pubmed.ncbi.nlm.nih.gov/31090027/)]
9. Tiwari A, Chaudhari M, Rai A. Multidisciplinary approach of artificial intelligence over medical imaging: a review, challenges, recent opportunities for research. In: Proceedings of the 3rd International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud). 2019 Presented at: I-SMAC '19; December 12-14, 2019; Palladam, India p. 237-242 URL: <https://ieeexplore.ieee.org/document/9032566> [doi: [10.1109/i-smac47947.2019.9032566](https://doi.org/10.1109/i-smac47947.2019.9032566)]
10. Cao J, Yao J. Linking different artificial intelligence functions to employees' psychological appraisals and work. *Acad Manag Proc* 2020 Aug;2020(1):19876. [doi: [10.5465/AMBPP.2020.105](https://doi.org/10.5465/AMBPP.2020.105)]
11. Gillan C, Milne E, Harnett N, Purdie TG, Jaffray DA, Hodges B. Professional implications of introducing artificial intelligence in healthcare: an evaluation using radiation medicine as a testing ground. *J Radiother Pract* 2018 Oct 03;18(1):5-9. [doi: [10.1017/s1460396918000468](https://doi.org/10.1017/s1460396918000468)]
12. Tekkeşin A. Artificial intelligence in healthcare: past, present and future. *Anatol J Cardiol* 2019 Oct;22(Suppl 2):8-9 [FREE Full text] [doi: [10.14744/AnatolJCardiol.2019.28661](https://doi.org/10.14744/AnatolJCardiol.2019.28661)] [Medline: [31670713](https://pubmed.ncbi.nlm.nih.gov/31670713/)]
13. Le Nguyen T. Blockchain in healthcare: a new technology benefit for both patients and doctors. In: Proceedings of the 2018 Portland International Conference on Management of Engineering and Technology. 2018 Presented at: PICMET '18; August 19-23, 2018; Honolulu, HI p. 1-6 URL: <https://ieeexplore.ieee.org/document/8481969> [doi: [10.23919/picmet.2018.8481969](https://doi.org/10.23919/picmet.2018.8481969)]
14. Hamet P, Tremblay J. Artificial intelligence in medicine. *Metabolism* 2017 Apr;69S:S36-S40. [doi: [10.1016/j.metabol.2017.01.011](https://doi.org/10.1016/j.metabol.2017.01.011)] [Medline: [28126242](https://pubmed.ncbi.nlm.nih.gov/28126242/)]
15. Mistry P. Artificial intelligence in primary care. *Br J Gen Pract* 2019 Sep;69(686):422-423 [FREE Full text] [doi: [10.3399/bjgp19X705137](https://doi.org/10.3399/bjgp19X705137)] [Medline: [31467001](https://pubmed.ncbi.nlm.nih.gov/31467001/)]
16. Yashchenko V. Artificial intelligence theory (basic concepts). In: Proceedings of the 2014 Science and Information Conference. 2014 Presented at: SAI '14; August 27-29, 2014; London, UK p. 473-480 URL: <https://ieeexplore.ieee.org/abstract/document/6918230> [doi: [10.1109/sai.2014.6918230](https://doi.org/10.1109/sai.2014.6918230)]
17. Jiang F, Jiang Y, Zhi H, Dong Y, Li H, Ma S, et al. Artificial intelligence in healthcare: past, present and future. *Stroke Vasc Neurol* 2017 Dec;2(4):230-243 [FREE Full text] [doi: [10.1136/svn-2017-000101](https://doi.org/10.1136/svn-2017-000101)] [Medline: [29507784](https://pubmed.ncbi.nlm.nih.gov/29507784/)]
18. Liyanage H, Liaw ST, Jonnagaddala J, Schreiber R, Kuziemsy C, Terry AL, et al. Artificial intelligence in primary health care: perceptions, issues, and challenges. *Yearb Med Inform* 2019 Aug;28(1):41-46 [FREE Full text] [doi: [10.1055/s-0039-1677901](https://doi.org/10.1055/s-0039-1677901)] [Medline: [31022751](https://pubmed.ncbi.nlm.nih.gov/31022751/)]
19. Abdullah R, Fakieh B. Health care employees' perceptions of the use of artificial intelligence applications: survey study. *J Med Internet Res* 2020 May 14;22(5):e17620 [FREE Full text] [doi: [10.2196/17620](https://doi.org/10.2196/17620)] [Medline: [32406857](https://pubmed.ncbi.nlm.nih.gov/32406857/)]
20. Kansal R, Bawa A, Bansal A, Trehan S, Goyal K, Goyal N, et al. Differences in knowledge and perspectives on the usage of artificial intelligence among doctors and medical students of a developing country: a cross-sectional study. *Cureus* 2022 Jan;14(1):e21434 [FREE Full text] [doi: [10.7759/cureus.21434](https://doi.org/10.7759/cureus.21434)] [Medline: [35223222](https://pubmed.ncbi.nlm.nih.gov/35223222/)]
21. Amisha, Malik P, Pathania M, Rathaur VK. Overview of artificial intelligence in medicine. *J Family Med Prim Care* 2019 Jul;8(7):2328-2331 [FREE Full text] [doi: [10.4103/jfmpc.jfmpc_440_19](https://doi.org/10.4103/jfmpc.jfmpc_440_19)] [Medline: [31463251](https://pubmed.ncbi.nlm.nih.gov/31463251/)]
22. Ahuja AS. The impact of artificial intelligence in medicine on the future role of the physician. *PeerJ* 2019;7:e7702 [FREE Full text] [doi: [10.7717/peerj.7702](https://doi.org/10.7717/peerj.7702)] [Medline: [31592346](https://pubmed.ncbi.nlm.nih.gov/31592346/)]
23. Davenport T, Kalakota R. The potential for artificial intelligence in healthcare. *Future Healthc J* 2019 Jun;6(2):94-98 [FREE Full text] [doi: [10.7861/futurehosp.6-2-94](https://doi.org/10.7861/futurehosp.6-2-94)] [Medline: [31363513](https://pubmed.ncbi.nlm.nih.gov/31363513/)]
24. Hamid S. The opportunities and risks of artificial intelligence in medicine and healthcare. *SPE Communications*. 2016. URL: https://www.cuspe.org/wp-content/uploads/2016/09/Hamid_2016.pdf [accessed 2024-03-23]

25. Keyworth C, Epton T, Goldthorpe J, Calam R, Armitage CJ. Are healthcare professionals delivering opportunistic behaviour change interventions? A multi-professional survey of engagement with public health policy. *Implement Sci* 2018 Sep 21;13(1):122 [FREE Full text] [doi: [10.1186/s13012-018-0814-x](https://doi.org/10.1186/s13012-018-0814-x)] [Medline: [30241557](https://pubmed.ncbi.nlm.nih.gov/30241557/)]
26. Scheetz J, Koca D, McGuinness M, Holloway E, Tan Z, Zhu Z, et al. Real-world artificial intelligence-based opportunistic screening for diabetic retinopathy in endocrinology and indigenous healthcare settings in Australia. *Sci Rep* 2021 Aug 04;11(1):15808 [FREE Full text] [doi: [10.1038/s41598-021-94178-5](https://doi.org/10.1038/s41598-021-94178-5)] [Medline: [34349130](https://pubmed.ncbi.nlm.nih.gov/34349130/)]
27. Tariq A, Patel BN, Sensakovic WF, Fahrenholtz SJ, Banerjee I. Opportunistic screening for low bone density using abdominopelvic computed tomography scans. *Med Phys* 2023 Jul;50(7):4296-4307. [doi: [10.1002/mp.16230](https://doi.org/10.1002/mp.16230)] [Medline: [36748265](https://pubmed.ncbi.nlm.nih.gov/36748265/)]
28. Bjerkén LV, Rønborg SN, Jensen MT, Ørting SN, Nielsen OW. Artificial intelligence enabled ECG screening for left ventricular systolic dysfunction: a systematic review. *Heart Fail Rev* 2022 Nov 08;28(2):419-430 [FREE Full text] [doi: [10.1007/s10741-022-10283-1](https://doi.org/10.1007/s10741-022-10283-1)] [Medline: [36344908](https://pubmed.ncbi.nlm.nih.gov/36344908/)]
29. Wang L, Zhang Y, Wang D, Tong X, Liu T, Zhang S, et al. Artificial intelligence for COVID-19: a systematic review. *Front Med (Lausanne)* 2021;8:704256 [FREE Full text] [doi: [10.3389/fmed.2021.704256](https://doi.org/10.3389/fmed.2021.704256)] [Medline: [34660623](https://pubmed.ncbi.nlm.nih.gov/34660623/)]
30. Kueper JK, Terry AL, Zwarenstein M, Lizotte DJ. Artificial intelligence and primary care research: a scoping review. *Ann Fam Med* 2020 May;18(3):250-258 [FREE Full text] [doi: [10.1370/afm.2518](https://doi.org/10.1370/afm.2518)] [Medline: [32393561](https://pubmed.ncbi.nlm.nih.gov/32393561/)]
31. Wang Y, Kung L, Wang WY, Cegielski CG. An integrated big data analytics-enabled transformation model: application to health care. *Inf Manag* 2018 Jan;55(1):64-79. [doi: [10.1016/j.im.2017.04.001](https://doi.org/10.1016/j.im.2017.04.001)]
32. Sutton RT, Pincock D, Baumgart DC, Sadowski DC, Fedorak RN, Kroeker KI. An overview of clinical decision support systems: benefits, risks, and strategies for success. *NPJ Digit Med* 2020;3:17 [FREE Full text] [doi: [10.1038/s41746-020-0221-y](https://doi.org/10.1038/s41746-020-0221-y)] [Medline: [32047862](https://pubmed.ncbi.nlm.nih.gov/32047862/)]
33. Dogru AK, Keskin BB. AI in operations management: applications, challenges and opportunities. *J Data Inf Manag* 2020 Feb 21;2(2):67-74. [doi: [10.1007/s42488-020-00023-1](https://doi.org/10.1007/s42488-020-00023-1)]
34. Jin Y, Yang H, Ji W, Wu W, Chen S, Zhang W, et al. Virology, epidemiology, pathogenesis, and control of COVID-19. *Viruses* 2020 Mar 27;12(4):372 [FREE Full text] [doi: [10.3390/v12040372](https://doi.org/10.3390/v12040372)] [Medline: [32230900](https://pubmed.ncbi.nlm.nih.gov/32230900/)]
35. Sarwar S, Dent A, Faust K, Richer M, Djuric U, Van Ommeren R, et al. Physician perspectives on integration of artificial intelligence into diagnostic pathology. *NPJ Digit Med* 2019;2:28 [FREE Full text] [doi: [10.1038/s41746-019-0106-0](https://doi.org/10.1038/s41746-019-0106-0)] [Medline: [31304375](https://pubmed.ncbi.nlm.nih.gov/31304375/)]
36. Moxley-Wyles B, Colling R, Verrill C. Artificial intelligence in pathology: an overview. *Diagn Histopathol* 2020 Nov;26(11):513-520 [FREE Full text] [doi: [10.1016/j.mpdhp.2020.08.004](https://doi.org/10.1016/j.mpdhp.2020.08.004)]
37. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* 2017 Feb 02;542(7639):115-118 [FREE Full text] [doi: [10.1038/nature21056](https://doi.org/10.1038/nature21056)] [Medline: [28117445](https://pubmed.ncbi.nlm.nih.gov/28117445/)]
38. Karches KE. Against the iDoctor: why artificial intelligence should not replace physician judgment. *Theor Med Bioeth* 2018 Apr;39(2):91-110. [doi: [10.1007/s11017-018-9442-3](https://doi.org/10.1007/s11017-018-9442-3)] [Medline: [29992371](https://pubmed.ncbi.nlm.nih.gov/29992371/)]
39. Meskó B, Hetényi G, Gyórfy Z. Will artificial intelligence solve the human resource crisis in healthcare? *BMC Health Serv Res* 2018 Jul 13;18(1):545 [FREE Full text] [doi: [10.1186/s12913-018-3359-4](https://doi.org/10.1186/s12913-018-3359-4)] [Medline: [30001717](https://pubmed.ncbi.nlm.nih.gov/30001717/)]
40. Farhud DD, Zokaei S. Ethical issues of artificial intelligence in medicine and healthcare. *Iran J Public Health* 2021 Nov;50(11):i-v [FREE Full text] [doi: [10.18502/ijph.v50i11.7600](https://doi.org/10.18502/ijph.v50i11.7600)] [Medline: [35223619](https://pubmed.ncbi.nlm.nih.gov/35223619/)]
41. Kerasidou A. Artificial intelligence and the ongoing need for empathy, compassion and trust in healthcare. *Bull World Health Organ* 2020 Apr 01;98(4):245-250 [FREE Full text] [doi: [10.2471/BLT.19.237198](https://doi.org/10.2471/BLT.19.237198)] [Medline: [32284647](https://pubmed.ncbi.nlm.nih.gov/32284647/)]
42. Zehir C, Karaboğa T, Başar D. The transformation of human resource management and its impact on overall business performance: big data analytics and AI technologies in strategic HRM. In: Hacioglu U, editor. *Digital Business Strategies in Blockchain Ecosystems: Transformational Design and Future of Global Business*. Cham, Switzerland: Springer; 2020:265-279.
43. Bhardwaj G, Singh SV, Kumar V. An empirical study of artificial intelligence and its impact on human resource functions. In: *Proceedings of the 2020 International Conference on Computation, Automation and Knowledge Management*. 2020 Presented at: ICCAKM '20; October 19-23, 2020; Dubai, United Arab Emirates p. 47-51 URL: <https://ieeexplore.ieee.org/document/9051544> [doi: [10.1109/iccakm46823.2020.9051544](https://doi.org/10.1109/iccakm46823.2020.9051544)]
44. Wolff J, Pauling J, Keck A, Baumbach J. The economic impact of artificial intelligence in health care: systematic review. *J Med Internet Res* 2020 Feb 20;22(2):e16866 [FREE Full text] [doi: [10.2196/16866](https://doi.org/10.2196/16866)] [Medline: [32130134](https://pubmed.ncbi.nlm.nih.gov/32130134/)]
45. Storey J. John Storey (ed.): *human resource management. A critical text*. *Organ Stud* 2016 Jul 01;17(1):158. [doi: [10.1177/017084069601700115](https://doi.org/10.1177/017084069601700115)]
46. Qiu L, Zhao L. Opportunities and challenges of artificial intelligence to human resource management. *Acad J Humanit Soc Sci* 2019;2(1):144-153 [FREE Full text] [doi: [10.25236/AJHSS.040036](https://doi.org/10.25236/AJHSS.040036)]
47. Irshaidat R. Interpretivism vs. positivism in political marketing research. *J Polit Mark* 2019 Jun 10;21(2):126-160. [doi: [10.1080/15377857.2019.1624286](https://doi.org/10.1080/15377857.2019.1624286)]
48. Mishra P, Pandey CM, Singh U, Gupta A, Sahu C, Keshri A. Descriptive statistics and normality tests for statistical data. *Ann Card Anaesth* 2019;22(1):67-72 [FREE Full text] [doi: [10.4103/aca.ACA_157_18](https://doi.org/10.4103/aca.ACA_157_18)] [Medline: [30648682](https://pubmed.ncbi.nlm.nih.gov/30648682/)]

49. Mukaka MM. Statistics corner: a guide to appropriate use of correlation coefficient in medical research. *Malawi Med J* 2012 Sep;24(3):69-71 [FREE Full text] [Medline: [23638278](#)]
50. Bolboacă SD, Jäntschi L, Sestraş AF, Sestraş RE, Pamfil DC. Pearson-fisher chi-square statistic revisited. *Information* 2011 Sep 15;2(3):528-545. [doi: [10.3390/info2030528](#)]
51. Franke TM, Ho T, Christie CA. The Chi-Square test. *Am J Eval* 2011 Nov 08;33(3):448-458. [doi: [10.1177/1098214011426594](#)]
52. Nahm FS. Nonparametric statistical tests for the continuous data: the basic concept and the practical use. *Korean J Anesthesiol* 2016 Feb;69(1):8-14 [FREE Full text] [doi: [10.4097/kjae.2016.69.1.8](#)] [Medline: [26885295](#)]
53. Eghbali-Babadi M, Feizi A, Khosravi A, Nouri F, Taheri M, Sarrafzadegan N. Development and evaluation of the psychometric properties of a hypertension self-care questionnaire. *ARYA Atheroscler* 2019 Sep;15(5):241-249 [FREE Full text] [doi: [10.22122/arya.v15i5.1835](#)] [Medline: [31949451](#)]
54. Ayatollahi H, Sarabi FZ, Langarizadeh M. Clinicians' knowledge and perception of telemedicine technology. *Perspect Health Inf Manag* 2015;12(Fall):1c [FREE Full text] [Medline: [26604872](#)]
55. Oh S, Kim JH, Choi SW, Lee HJ, Hong J, Kwon SH. Physician confidence in artificial intelligence: an online mobile survey. *J Med Internet Res* 2019 Mar 25;21(3):e12422 [FREE Full text] [doi: [10.2196/12422](#)] [Medline: [30907742](#)]
56. Summerton N, Cansdale M. Artificial intelligence and diagnosis in general practice. *Br J Gen Pract* 2019 Jun 27;69(684):324-325. [doi: [10.3399/bjgp19x704165](#)]
57. Johansson J, Herranen S. The application of artificial intelligence (AI) in human resource management: current state of AI and its impact on the traditional recruitment process. Jönköping University. 2019 May. URL: <https://www.diva-portal.org/smash/get/diva2:1322478/FULLTEXT01.pdf> [accessed 2024-03-23]
58. Jia Q, Guo Y, Li R, Li Y, Chen Y. A conceptual artificial intelligence application framework in human resource management. In: Proceedings of the 18th International Conference on Electronic Business. 2018 Presented at: ICEB '18; December 2-6, 2018; Guilin, China p. 106-114 URL: <https://aisel.aisnet.org/iceb2018/91/>
59. Lee D, Yoon SN. Application of artificial intelligence-based technologies in the healthcare industry: opportunities and challenges. *Int J Environ Res Public Health* 2021 Jan 01;18(1):271 [FREE Full text] [doi: [10.3390/ijerph18010271](#)] [Medline: [33401373](#)]
60. Tran VT, Riveros C, Ravaud P. Patients' views of wearable devices and AI in healthcare: findings from the ComPaRe e-cohort. *NPJ Digit Med* 2019;2:53 [FREE Full text] [doi: [10.1038/s41746-019-0132-y](#)] [Medline: [31304399](#)]
61. Hazarika I. Artificial intelligence: opportunities and implications for the health workforce. *Int Health* 2020 Jul 01;12(4):241-245 [FREE Full text] [doi: [10.1093/inthealth/ihaa007](#)] [Medline: [32300794](#)]
62. Arora N, Banerjee AK, Narasu ML. The role of artificial intelligence in tackling COVID-19. *Future Virol* 2020 Nov;15(11):717-724. [doi: [10.2217/fvl-2020-0130](#)]
63. Lalmuanawma S, Hussain J, Chhakchhuak L. Applications of machine learning and artificial intelligence for COVID-19 (SARS-CoV-2) pandemic: a review. *Chaos Solit Fractals* 2020 Oct;139:110059 [FREE Full text] [doi: [10.1016/j.chaos.2020.110059](#)] [Medline: [32834612](#)]
64. Kolbjørnsrud V, Amico R, Thomas RJ. How artificial intelligence will redefine management. *Harvard business review*. 2019 Jun 23. URL: <https://hbr.org/2016/11/how-artificial-intelligence-will-redefine-management> [accessed 2024-03-23]
65. Apaydin E. Administrative work and job role beliefs in primary care physicians: an analysis of semi-structured interviews. *SAGE Open* 2020 Jan 09;10(1):215824401989909. [doi: [10.1177/2158244019899092](#)]
66. Gerke S, Minssen T, Cohen IG. Ethical and legal challenges of artificial intelligence-driven health care. In: Bohr A, Memarzadeh K, editors. *Artificial Intelligence in Healthcare*. Cambridge, MA: Academic Press; 2020.
67. Nagy M, Sisk B. How will artificial intelligence affect patient-clinician relationships? *AMA J Ethics* 2020 May 01;22(5):E395-E400 [FREE Full text] [doi: [10.1001/amajethics.2020.395](#)] [Medline: [32449655](#)]
68. Ross PT, Bibler Zaidi NL. Limited by our limitations. *Perspect Med Educ* 2019 Aug;8(4):261-264 [FREE Full text] [doi: [10.1007/s40037-019-00530-x](#)] [Medline: [31347033](#)]
69. Qumseya B, Goddard A, Qumseya A, Estores D, Draganov PV, Forsmark C. Barriers to clinical practice guideline implementation among physicians: a physician survey. *Int J Gen Med* 2021;14:7591-7598 [FREE Full text] [doi: [10.2147/IJGM.S333501](#)] [Medline: [34754231](#)]
70. Delnevo CD, Singh B. The effect of a web-push survey on physician survey responses rates: a randomized experiment. *Surv Pract* 2021;14(1) [FREE Full text] [doi: [10.29115/sp-2021-0001](#)] [Medline: [33604202](#)]
71. Creswell JW, Fetters MD, Ivankova NV. Designing a mixed methods study in primary care. *Ann Fam Med* 2004;2(1):7-12 [FREE Full text] [doi: [10.1370/afm.104](#)] [Medline: [15053277](#)]

Abbreviations

- AI:** artificial intelligence
- CT:** computed tomography
- DL:** deep learning
- ML:** machine learning

PHCC: Primary Health Care Corporation

Edited by K El Emam; submitted 06.07.22; peer-reviewed by D Paradice, L Novak, R Sánchez de Madariaga; comments to author 30.01.23; revised version received 25.05.23; accepted 07.03.24; published 17.04.24.

Please cite as:

Waheed MA, Liu L

Perceptions of Family Physicians About Applying AI in Primary Health Care: Case Study From a Premier Health Care Organization
JMIR AI 2024;3:e40781

URL: <https://ai.jmir.org/2024/1/e40781>

doi: [10.2196/40781](https://doi.org/10.2196/40781)

PMID: [38875531](https://pubmed.ncbi.nlm.nih.gov/38875531/)

©Muhammad Atif Waheed, Lu Liu. Originally published in JMIR AI (<https://ai.jmir.org>), 17.04.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR AI, is properly cited. The complete bibliographic information, a link to the original publication on <https://www.ai.jmir.org/>, as well as this copyright and license information must be included.

Original Paper

Feasibility of Multimodal Artificial Intelligence Using GPT-4 Vision for the Classification of Middle Ear Disease: Qualitative Study and Validation

Masao Noda^{1*}, MBA, MD, PhD; Hidekane Yoshimura^{2*}, MD, PhD; Takuya Okubo², MD; Ryota Kosu¹, MD; Yuki Uchiyama², MD; Akihiro Nomura³, MD, PhD; Makoto Ito¹, MD, PhD; Yutaka Takumi², MD, PhD

¹Department of Otolaryngology, Head and Neck Surgery, Jichi Medical University, Shimotsuke, Japan

²Department of Otolaryngology - Head and Neck Surgery, Shinshu University, Matsumoto, Japan

³College of Transdisciplinary Sciences for Innovation, Kanazawa University, Kanazawa, Japan

*these authors contributed equally

Corresponding Author:

Masao Noda, MBA, MD, PhD

Department of Otolaryngology, Head and Neck Surgery

Jichi Medical University

3311-1 Yakushiji

Shimotsuke, 329-0498

Japan

Phone: 81 285442111

Email: doforanabdosuc@gmail.com

Related Article:

This is a corrected version. See correction statement: <https://ai.jmir.org/2024/1/e62990/>

Abstract

Background: The integration of artificial intelligence (AI), particularly deep learning models, has transformed the landscape of medical technology, especially in the field of diagnosis using imaging and physiological data. In otolaryngology, AI has shown promise in image classification for middle ear diseases. However, existing models often lack patient-specific data and clinical context, limiting their universal applicability. The emergence of GPT-4 Vision (GPT-4V) has enabled a multimodal diagnostic approach, integrating language processing with image analysis.

Objective: In this study, we investigated the effectiveness of GPT-4V in diagnosing middle ear diseases by integrating patient-specific data with otoscopic images of the tympanic membrane.

Methods: The design of this study was divided into two phases: (1) establishing a model with appropriate prompts and (2) validating the ability of the optimal prompt model to classify images. In total, 305 otoscopic images of 4 middle ear diseases (acute otitis media, middle ear cholesteatoma, chronic otitis media, and otitis media with effusion) were obtained from patients who visited Shinshu University or Jichi Medical University between April 2010 and December 2023. The optimized GPT-4V settings were established using prompts and patients' data, and the model created with the optimal prompt was used to verify the diagnostic accuracy of GPT-4V on 190 images. To compare the diagnostic accuracy of GPT-4V with that of physicians, 30 clinicians completed a web-based questionnaire consisting of 190 images.

Results: The multimodal AI approach achieved an accuracy of 82.1%, which is superior to that of certified pediatricians at 70.6%, but trailing behind that of otolaryngologists at more than 95%. The model's disease-specific accuracy rates were 89.2% for acute otitis media, 76.5% for chronic otitis media, 79.3% for middle ear cholesteatoma, and 85.7% for otitis media with effusion, which highlights the need for disease-specific optimization. Comparisons with physicians revealed promising results, suggesting the potential of GPT-4V to augment clinical decision-making.

Conclusions: Despite its advantages, challenges such as data privacy and ethical considerations must be addressed. Overall, this study underscores the potential of multimodal AI for enhancing diagnostic accuracy and improving patient care in otolaryngology. Further research is warranted to optimize and validate this approach in diverse clinical settings.

KEYWORDS

artificial intelligence; deep learning; machine learning; generative AI; generative; tympanic membrane; middle ear disease; GPT4-Vision; otolaryngology; ears; ear; tympanic; vision; GPT; GPT4V; otoscopic; image; images; imaging; diagnosis; diagnoses; diagnostic; diagnostics; otitis; mobile phone

Introduction

The emergence of artificial intelligence (AI) has altered the landscape of medical technology, particularly in diagnosis, which leverages the identification of features based on imaging and physiological data [1-3]. In the field of otolaryngology, AI and deep learning models are being used for imaging; ongoing efforts focus on classifying diseases based on tympanic membrane images of middle ear disease [4-6]. Technological advancements, including deep learning and transfer learning using pretrained models, have resulted in an accuracy range of 70%-90% in models for analyzing otoscopic images [7]. There have also been advancements in its application, such as implementing smartphone-based point-of-care diagnostics [8]. However, these models rely on trained image data, require large image data sets, and do not consider patient information or clinical context. Consequently, the universality of these models is limited, and their optimal application in clinical practice remains unclear.

Recently, large-scale language-processing models have become available for general use. Further, 1 such model, the GPT-4, has demonstrated specialist-level medical knowledge through its language-processing abilities [9-11]. Since October 2023, GPT-4 Vision (GPT-4V) has gained the ability to evaluate image data, enabling a multimodal diagnostic approach that incorporates both language processing and image analysis [12]. GPT-4V enables the integration of patient information analysis and image-based deep learning models, providing valuable

support in diagnosis and treatment, similar to decisions made in a clinical setting [13]. Multimodal AI, which bases diagnosis on multiple pieces of information, has been reported to be more effective than methods that rely on a single type of information. This is demonstrated in various medical applications, including the combination of pathology images with genomic information [14] and their use in liver cancer [15] and cervical cancer [16], where imaging information is integrated. In otorhinolaryngology, there have been few reports; however, efforts to incorporate AI for otoscopic images could further improve the quality of care.

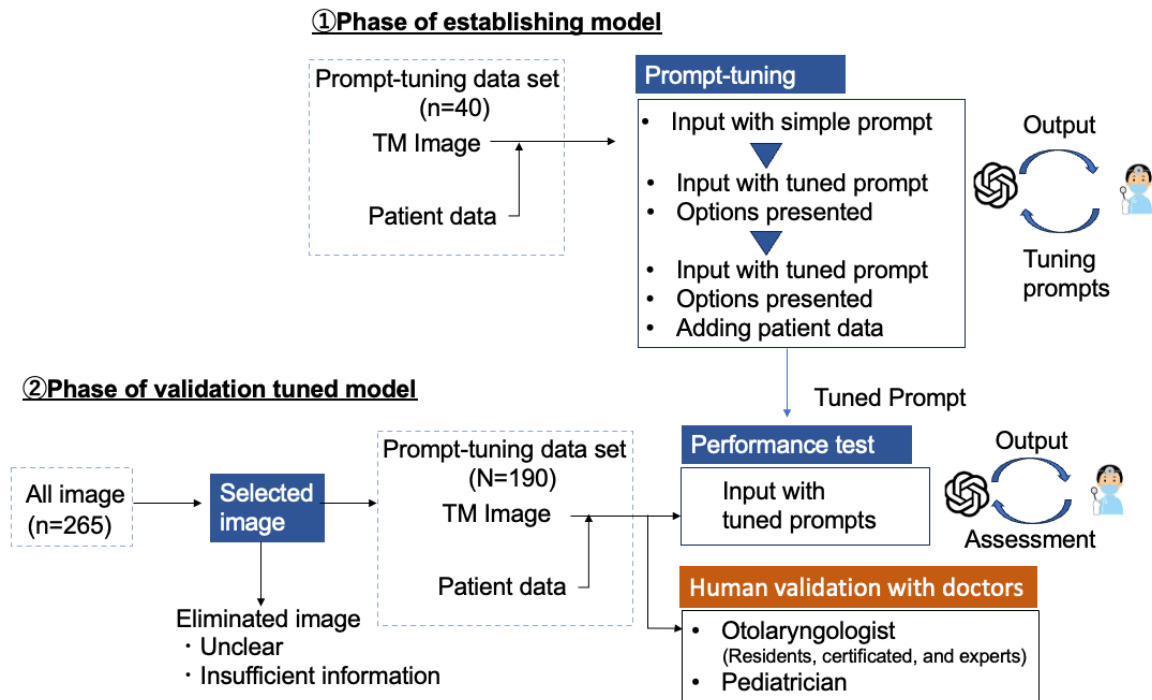
In this study, we aimed to investigate the effectiveness of a multimodal approach using GPT-4V to diagnose middle ear disease. This approach was designed to integrate patient-specific data (age, sex, and chief complaint) with tympanic membrane images to assess the accuracy of the versatile GPT-4V. The model's accuracy was compared with physicians' diagnoses to validate its effectiveness in image-based deep learning. The potential future development of the multimodal AI approach for classifying middle ear diseases is also discussed.

Methods

Study Design

GPT-4V has been available as an image recognition model since September 25, 2023. This study's design was divided into two phases: (1) establishing a model with appropriate prompts and (2) validating the ability of the optimal prompt model to classify images (Figure 1).

Figure 1. Overview of this study. The model was divided into two phases: (1) establishment and (2) tuned model validation. TM: tympanic membrane.



Correct Otosopic Images and Patient Information

This study included 305 otoscopic images of middle ear disease obtained from patients who visited Shinshu University or Jichi Medical University between April 2010 and December 2023. The endoscope used was an Olympus ENF-VH and ENF-V3 (Olympus), and the video system was an Olympus VISERA ELITE OTV-S190. Further, 1 image was obtained from each patient. We excluded images with poor quality and those in which multiple diseases were suspected. The remaining images were classified into 4 disease categories: acute otitis media (AOM), middle ear cholesteatoma (chole), chronic otitis media (COM), and otitis media with effusion (OME). The final diagnoses were based on the judgment of the otolaryngologists who treated the patients. These images were accompanied by patient-specific information, such as age, sex, and chief complaint (eg, fever, otalgia, otorrhea, ear fullness, deafness, facial palsy, dizziness, and tinnitus). We excluded images taken after otologic surgery. Of note, only 1 image was obtained from each patient.

GPT-4V Settings and Prompt Tuning

The GPT-4V settings were established using prompts reported in previous studies [17,18]. Briefly, conditions and prompts for providing answers were verified using 10 images for each disease. According to a report on prompts [19], image data or patient information were manually input into GPT-4V, and the generated results were evaluated by the physicians (MN and HY).

Accuracy Verification of GPT-4V Using the Optimal Prompt Model

The model with the optimal prompt created was used to verify the diagnostic accuracy of GPT-4V on 190 images (37 in AOM, 53 in chole [6 in congenital, 47 in acquired], 51 in COM, and

49 in OME), which were different from those for tuning prompts. To account for the variability in responses, each administration was performed 3 times, and responses that were answered 2 or more times were considered to be the actual response.

Comparison of AI Accuracy With Physician Accuracy

To compare the diagnostic accuracy of GPT-4V with that of physicians, 30 clinicians completed a web-based questionnaire consisting of 190 images.

The web-based survey included tympanic membrane images and patient information (age, sex, and chief complaint) in a 4-choice question format. The respondents included 8 certificated pediatricians, 8 otolaryngology residents, 8 certificated otolaryngologists, and 6 experts in otolaryngology (more than 15 years of experience).

To show the trend in the percentage of correct responses according to the difficulty of the questions, the questions were divided into 3 levels (easy, normal, and hard) according to the overall percentage of correct responses by physicians, and the percentage of correct responses for each level and each question was compared between the GPT-4V and all doctors, otolaryngologists, and pediatricians.

Ethical Considerations

Patient information was anonymized to protect privacy and used only with the approval of the Ethics Committee of the Shinshu University School of Medicine (6088).

Statistical Analysis

Groups were compared by 1-way ANOVA. Subsequently, multiple comparison tests (the Bonferroni method) were used to compare groups. Statistical significance was set at $P < .05$. A 1-sample proportion test was used to compare the performance

of the physician with that of GPT-4V in terms of the correct response rate.

Results

Establishment of Optimal Prompts

In the initial stage, we sought an optimal input method using 10 images for each disease (AOM, chole, COM, or OME; 40 images total). First, we input only images or options; GPT mostly requires clinical information, such as patient history and symptoms, although no response regarding the disease was generated (Figure 1 and Multimedia Appendix 1). Second, the names of the 4 diseases were added as candidate answers, but again, no response regarding the disease was generated. When detailed patient information, such as age, sex, and main symptoms, was inputted, GPT-4V provided answers, indicating that input images with patient data were the optimal prompt for testing the accuracy of GPT-4V.

Accuracy Validation of the Multimodal AI Approach

The performance of the multimodal AI approach in this study for classifying middle ear diseases was validated, with an overall diagnostic accuracy of 82.1% for the GPT-4V-based analysis. Disease-specific accuracy rates were 89.2% for AOM (true positives [TP]=33, false positives [FP]=1, false negatives [FN]=4, precision=0.97, recall=0.89, F_1 -score=0.93), 76.5% for COM (TP=39, FP=7, FN=12, precision=0.85, recall=0.76, F_1 -score=0.8), 79.3% for cholesteatoma (TP=42, FP=13, FN=11, precision=0.76, recall=0.79, F_1 -score=0.78), and 85.7% for OME (TP=42, FP=10, FN=7, precision=0.81, recall=0.86, F_1 -score=0.83; Figure 2).

These results indicate high discrimination among various disease types; however, there were also some incorrect responses. Representative images of correct and incorrect GPT-4V classifications for each disease are shown in Figure 3.

Figure 2. Confusion matrix of GPT-4V for classifying 4 middle ear diseases. AOM: acute otitis media; chole: middle ear cholesteatoma; COM: chronic otitis media; GPT-4V: GPT-4 Vision; OME: otitis media with effusion.

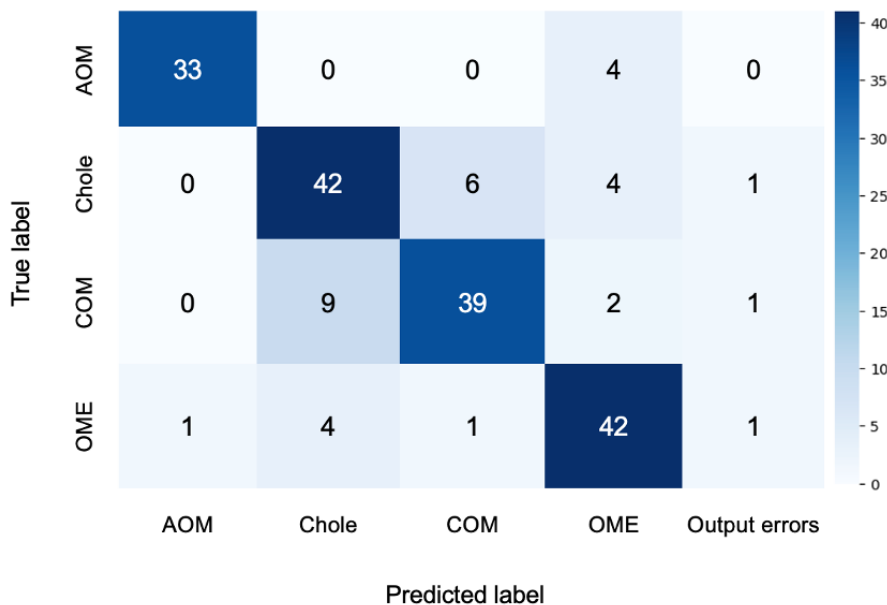
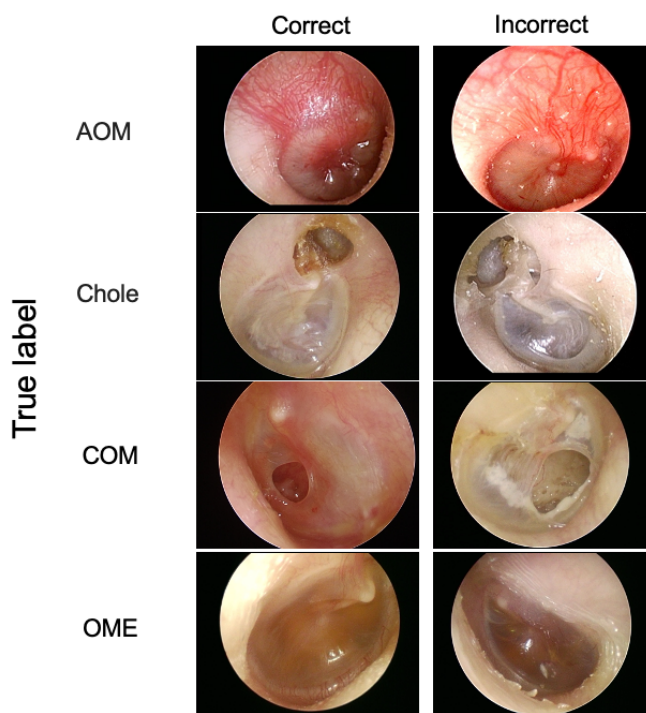


Figure 3. Representative images of correct and incorrect GPT-4V classifications for 4 middle ear diseases. The left side shows the correct images for GPT-4V classification, and the right side shows the incorrect images for GPT-4V. AOM: acute otitis media; chole: middle ear cholesteatoma; COM: chronic otitis media; GPT-4V: GPT-4 Vision; OME: otitis media with effusion.



Comparison of Diagnostic Accuracy by Physicians and GPT-4V

The same images with patients' information used by GPT-4V were evaluated by pediatricians (n=8), otolaryngology residents (n=8), certificated otolaryngologists (n=8), and experts in otolaryngology (n=6), and the diagnostic accuracy of each group was compared. The mean diagnostic accuracy was 70.6% (SE 4.2%) for pediatricians, 95.5% (SE 1%) for otolaryngology residents, 97.3% (SE 0.8%) for certificated otolaryngologists, and 98.2% (SE 0.4%) for experts in otolaryngology. ANOVA revealed significant differences among the 4 groups ($F_1=13.43$, $P<.001$). In the post hoc comparison, a significant difference was observed between pediatricians and the other 3 groups ($P<.001$). The GPT-4V correct response rate was 82.1%, surpassing that of pediatricians by 11.5% and trailing behind otolaryngologists by an average of just over 10% (Figure 4).

The accuracy rates for specific diseases were as follows: 92.3% for AOM (pediatricians 80.4%, otolaryngology residents 94.9%, certificated otolaryngologists 97%, and experts in otolaryngology 98.2%), 95.9% for COM (pediatricians 89.5%, otolaryngology residents 96.6%, certificated otolaryngologists 99.8%, and experts in otolaryngology 98.4%), 81.8% for chole (pediatricians 46%, otolaryngology residents 93.2%, certificated otolaryngologists 93.6%, and experts in otolaryngology 98.4%), and 93.7% for OME (pediatricians 81.6%, otolaryngology residents 97.2%, certificated otolaryngologists 99%, and experts in otolaryngology 98%).

In the confusion matrix of all doctors, there was a notable tendency to misclassify chole as OME and AOM as OME. Among pediatricians, there were more errors in classifying chole as AOM or COM (Figure 5).

Figure 4. Result of human validations with doctors of TM images and patients' data. The graph shows the average correct rate for doctors (pediatricians, otolaryngology residents, certificated otolaryngologists, and experts in otolaryngology), and the dotted line shows the correct answer rate of GPT-4V. GPT-4V: GPT-4 Vision; TM: tympanic membrane. ****P** value <.01.

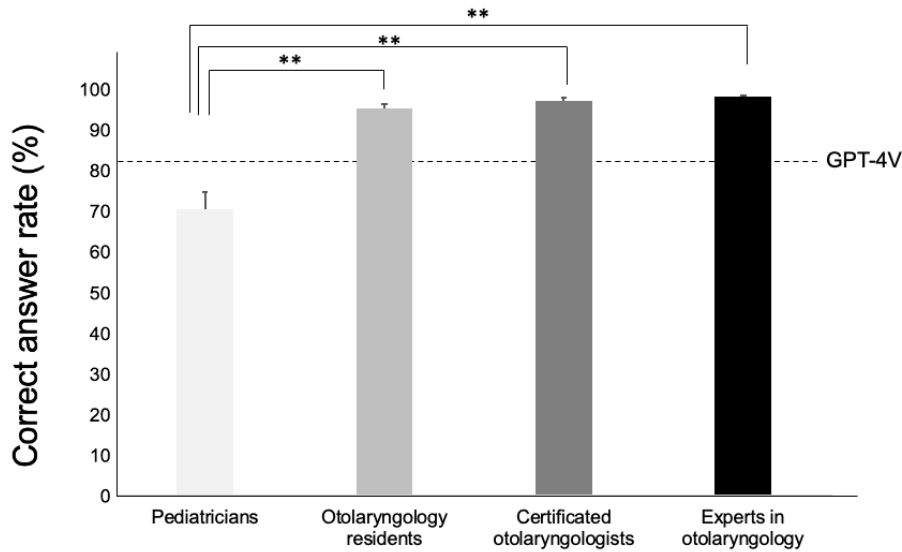
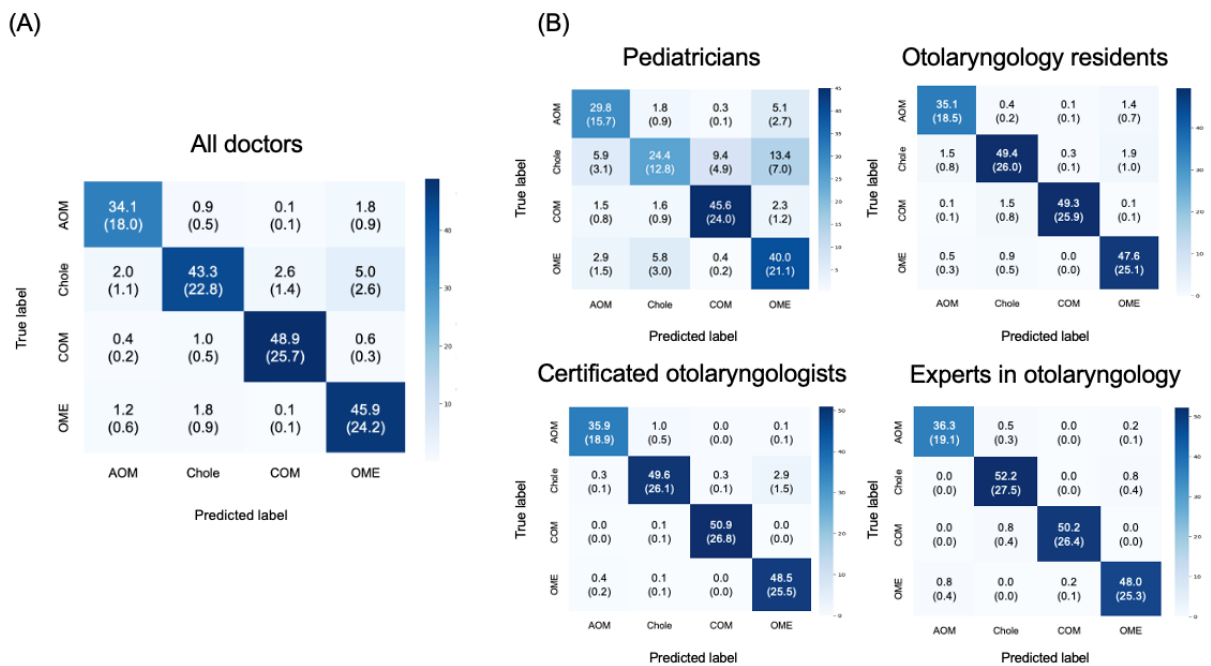


Figure 5. Confusion matrix of doctors (pediatricians, otolaryngology residents, certificated otolaryngologists, and experts in otolaryngology) for classifying 4 middle ear diseases. (A) Confusion matrix of all doctors (N=30). The average (percentage of total responses) is shown. (B) Confusion matrix of doctors in each group: pediatricians (n=8), otolaryngology residents (n=8), certificated otolaryngologists (n=8), and experts in otolaryngology (n=6). The averages of each group (percentage of total responses) are shown. AOM: acute otitis media; chole: cholesteatoma; COM: chronic otitis media; OME: otitis media with effusion.



Regarding the difference in the trend of the percentage of correct answers between GPT-4V and physicians according to the difficulty of the questions, even the percentage of correct answers for GPT-4V tended to decrease gradually from 85.7% for easy, 84% for normal, and 71.1% for hard questions (Table 1).

Furthermore, compared with otolaryngologists, GPT-4V had a significantly lower percentage of correct answers for all

questions (99.7% for easy, 97.1% for normal, and 90.8% for hard questions; all $P < .001$). In contrast, the results of the “hard” and “normal” groups were similar. Compared with pediatricians, the GPT-4V outperformed the pediatricians in easy questions with 96.6%, although no statistically significant difference was observed ($P = .006$). However, the GPT-4V had a predominantly higher percentage of correct answers for normal (76.3%, $P = .07$) and hard questions (45.4%, $P < .001$).

Table 1. Comparison of the scores by GPT-4 Vision (GPT-4V) and human validation with physicians across various difficulty levels (N=190).

Difficulty level	Questions, n (%)	GPT-4V (mean %)	All doctors			Otolaryngologists			Pediatricians		
			Mean % (95% CI)	Differences	P value	Mean % (95% CI)	Differences	P value	Mean % (95% CI)	Differences	P value
Easy (>95%)	77 (40.5)	85.7	97.8 (97.4-98.2)	12.1	<.001 ^a	99.7 (99.5-99.9)	14.0	<.001 ^a	96.6 (95.3-97.9)	10.9	.006 ^a
Normal (>85%, <95%)	75 (39.5)	84	90.4 (89.7-91.0)	6.4	.13	97.1 (96.2-98.0)	13.1	<.001 ^a	76.3 (73.6-79.0)	-7.7	.07
Hard (<85%)	38 (20)	71.1	76.8 (73.7-79.8)	5.7	.44	90.8 (87.2-94.3)	19.7	<.001 ^a	45.4 (39.5-51.3)	-25.7	<.001 ^a

^aStatistically significant.

Discussion

Principal Results

In this study, we assessed the accuracy of the GPT-4V multimodal AI approach in classifying middle ear disorders, yielding the following three key findings. First, GPT-4V, a general-purpose model focusing on large-scale language models, achieved approximately 80% accuracy in classifying middle ear disease. The model's performance, evaluated using images and patient data, was superior to that of nonotolaryngologists, although it was lower than the average accuracy of otolaryngologists. Second, the GPT-4V was able to classify diseases when patient information and disease options were input. Further improvements in accuracy could be achieved with more detailed patient information. Third, accuracy varied by disease, suggesting the potential for optimizing AI usage and improving accuracy by understanding the specificity of GPT-4V in classifying particular diseases.

Comparison With Prior Work

The GPT-4V model has undergone training and uses 0-shot learning, which recognizes image features based on natural language to classify diseases based on image information and previously learned disease features [20]. GPT-4V can yield effective results with fewer resources than previous deep learning models, which typically require a large amount of image data, computational resources, time, and parameter adjustments for training. By inputting new information rather than simply classifying image data, it becomes possible to tailor diagnoses and diagnostic aids for each individual. Furthermore, GPT-4V and other large-scale language processing models feature prompt development that is appropriate for its usage purposes, since the accuracy of such models varies depending on the prompt adjustments.

Compared with physicians' accuracy, the model's performance in this study was higher than that of a pediatrician but lower than that of an otolaryngologist. In a previous comparison between deep learning and humans, Crowson et al [21] classified 22 tympanic membrane images and found that the deep learning model achieved an accuracy of 95.5%, compared with an

accuracy of 65% for 39 clinicians. Suresh et al [22] also reported that a machine-learning model created from 1000 images was more effective than pediatricians, with an accuracy rate of 90.6%, surpassing the clinicians' accuracy of 59.4%. Our results indicated that the model did not reach the proficiency level of otolaryngologists; however, it could be valuable for using tympanic membrane images in medical practice outside of otolaryngology. In particular, GPT-4V judgments predominantly exceeded pediatricians' correct response rates for questions with normal to hard difficulty, suggesting that the present model may be useful for nonotolaryngologists who have difficulty in making such judgments.

Moreover, previous reports on deep learning classification models have determined the presence or absence of inflammation and exudates based on photographs alone. Further studies are needed to identify the optimal stage in the examination for implementing the image classification model and the subsequent policy decisions that should follow.

GPT-4V allows for the classification of diseases using patient information. While comments about medical or harmful content (with restrictions on medical advice) may result in a lower correct response rate, informative or educational responses are still possible if they are well-informed. Efforts have been made to use large language models (LLMs) to improve the accuracy of prompts. Therefore, it is possible to develop appropriate prompts for medical imaging and middle ear disorders. The accuracy of the LLM is expected to further improve with the development of prompts that are specifically tailored for medical imaging and middle ear disease [23,24].

For the clinical application of the GPT-4V model, collecting clinical data and adjusting parameters are needed to further improve its diagnostic accuracy for each middle ear disease. Upon reviewing the incorrect responses of GPT-4V for each disease, we found that chole might demonstrate a retraction pocket, which may be mistaken for a perforation. However, images with keratin debris accumulation in the retraction pocket were less prone to misclassification. In cases of COM with calcification, a white lesion was considered to be chole calcification, emphasizing the importance of distinguishing between these 2 diseases. AOM cases without the chief

complaint of acute inflammation (fever, ear pain, or ear discharge) were occasionally misclassified, even with characteristic findings such as a bulging tympanic membrane, suggesting that GPT-4V was likely to prioritize patients' information over images. In OME cases, a white lesion was sometimes considered to be a pearly tumor (chole) or tympanic membrane perforation (COM), particularly when it involved a small amount of effusion or air. For physicians, chole and AOM were often misidentified as other diseases and OME, respectively. When comparing the GPT-4V model with the entire group of physicians, the percentage of correct responses was generally higher among the physicians. However, the GPT-4V diagnostic accuracy for chole was higher than that of pediatricians, indicating that GPT-4V could help nonotolaryngologists diagnose chole. In a previous report, a dedicated AI model had a diagnostic accuracy of approximately 90% for chole [25]; therefore, the combination of such a system and GPT-4V would be useful to improve the accuracy of chole detection.

As demonstrated in this study, the application of AI, including LLM, is believed to offer advantages in terms of improving efficiency and providing assistance in clinical work, enabling the delivery of high-quality medical care, and overcoming language barriers in medicine. The use of GPT-4V has already been reported to diagnose complicated cases [26], and its application can be expanded by integrating it with imaging information. In the field of orthopedics, trials are underway to determine treatment methods based on MRI reports [27], showcasing the effectiveness of GPT-4V as an aid in image interpretation. GPT has been shown to return answers and provide details about the disease, including risk factors and treatment methods. This allows for the evaluation of images alone and assists in medical treatment. Such insights are valuable for understanding the practical use and challenges of AI in real-world applications. Unlike the simplistic deep learning models of the past, the LLM can enhance accuracy by presenting evidence for judgments and asking a series of questions. When used by physicians with a certain level of specialized knowledge, the LLM effectively aids judgment, leading to increased efficiency in medical care. GPT-4V provides answers in just a few seconds, which is significantly shorter than the time it takes a physician to provide a diagnosis, thereby confirming its efficiency. GPT-4V can be used on smartphones, potentially making medical treatment more location-independent. However, there are associated risks, including the reliance on AI for

medical care, misdiagnoses due to system malfunctions, and patient information leakage. ChatGPT (OpenAI, Microsoft Corporation) is trained based on information up to a certain period but may respond differently at different times or provide answers using outdated criteria. Furthermore, legal and personal literacy measures must be developed to protect personal information and address ethical concerns. Foreign countries and the United Nations are actively promoting laws and regulations governing the use of AI [28,29].

Limitations

In total, one limitation of this study is the use of a limited number of images (N=190). Further analysis is required to assess the impact of using a larger data set that encompasses various diseases. Additionally, as there are large variations in the quality of otoscopic images, accurate diagnosis might be challenging in some cases.

The recognition and content of the answers may change depending on the doctor, clinics, and designed prompt; the accuracy may also change due to changes in the image quality used or the method used to capture the image. While this is common to deep learning, the advantage of GPT, which does not require prior training, is that it is not affected by the data to be trained; thus, the possibility of such changes is considered to be small.

For these reasons, further exploration is needed on strategies for handling challenging images and facilitating open-ended responses without giving predefined options. Furthermore, because of the rapid pace of technological evolution, it is essential to regularly fine-tune and make a standalone model that ensures reliability and consistency over time.

Conclusions

A multimodal AI approach using GPT-4V has revealed a potential new diagnostic approach for classifying middle ear diseases. This confirms the ability of AI to assist in clinical diagnosis and identify disease-specific features. The significant improvement in accuracy compared with conventional deep learning models indicates that even general-purpose AI technology can assist in medical treatment with a certain level of accuracy. It can be applied to highly specialized diagnoses, depending on the method. Further improvements in diagnostic accuracy are expected in future studies by integrating more diverse data types.

Acknowledgments

We are thankful to our colleagues at Shinshu University: Dr Sota Ichimura, Dr Kota Hirose, Dr Mariko Kasuga, Dr Shu Yokota, Dr Kentaro Hori, Dr Arisa Oguchi, Dr Kenjiro Sugiyama, Dr Jun Shinagawa, Dr Yoichiro Iwasa, Dr Keita Tsukada, Dr Tomohiro Oguchi, and Dr Nobuyoshi Suzuki of the Department of Otolaryngology—Head and Neck, and those at Jichi Medical University: Dr Kota Matsuyama, Dr Akiko Uchida, and Dr Yuki Miura of the Department of Otolaryngology and Dr Shinya Fukuda, Dr Kazuki Okumura, and Dr Keizo Wakae of the Department of Pediatrics, and Dr Hitoshi Irabu, Dr Kazuhiro Noguchi, Dr Ryo Nakagawa, Dr Narutoshi Yamazaki, Dr Yuji Takaso, Dr Keisuke Koyama, Dr Yukari Nakamura, Dr Chia Sasaki, and Dr Keigo Nishida for their invaluable cooperation in this study. We thank Editage [30] for English language editing. The authors declare that no financial support was received for the research, authorship, or publication of this paper.

Authors' Contributions

MN handled the conceptualization, data curation, formal analysis, investigation, methodology, project administration, resources, validation, visualization, and writing of the original draft. HY worked on the conceptualization, data curation, formal analysis, investigation, methodology, project administration, resources, validation, and review and editing of the writing. TO, RK, and YU handled the investigation, project administration, and review and editing of the writing. AN, MI, and YT did the supervision and review and editing of the writing.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Representative image and prompt of this study. (A) Representative image of input and output to GPT-4 Vision. Input can be combined with text and images in input to obtain output. (B) Example of changing the prompt content and an output that asks for patient information. By presenting a concept as ORDER and adding conditions as restriction, appropriate prompts were attempted to be developed. In the output, it is required to input patient information such as age, medical history, and chief complaint. (C) An example of an answer with an optimized prompt. Present the diagnosis, the rationale for the diagnosis, and treatment and prevention methods.

[[PDF File \(Adobe PDF File\), 483 KB - ai_v3i1e58342_app1.pdf](#)]

References

1. Kermany DS, Goldbaum M, Cai W, Valentim CCS, Liang H, Baxter SL, et al. Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell* 2018;172(5):1122-1131.e9 [FREE Full text] [doi: [10.1016/j.cell.2018.02.010](https://doi.org/10.1016/j.cell.2018.02.010)] [Medline: [29474911](https://pubmed.ncbi.nlm.nih.gov/29474911/)]
2. Schaefferkoetter J, Yan J, Moon S, Chan R, Ortega C, Metser U, et al. Deep learning for whole-body medical image generation. *Eur J Nucl Med Mol Imaging* 2021;48(12):3817-3826 [FREE Full text] [doi: [10.1007/s00259-021-05413-0](https://doi.org/10.1007/s00259-021-05413-0)] [Medline: [34021779](https://pubmed.ncbi.nlm.nih.gov/34021779/)]
3. Lee CC, Lin CS, Tsai CS, Tsao TP, Cheng CC, Liou JT, et al. A deep learning-based system capable of detecting pneumothorax via electrocardiogram. *Eur J Trauma Emerg Surg* 2022;48(4):3317-3326 [FREE Full text] [doi: [10.1007/s00068-022-01904-3](https://doi.org/10.1007/s00068-022-01904-3)] [Medline: [35166869](https://pubmed.ncbi.nlm.nih.gov/35166869/)]
4. Choi Y, Chae J, Park K, Hur J, Kweon J, Ahn JH. Automated multi-class classification for prediction of tympanic membrane changes with deep learning models. *PLoS One* 2022;17(10):e0275846 [FREE Full text] [doi: [10.1371/journal.pone.0275846](https://doi.org/10.1371/journal.pone.0275846)] [Medline: [36215265](https://pubmed.ncbi.nlm.nih.gov/36215265/)]
5. Park YS, Jeon JH, Kong TH, Chung TY, Seo YJ. Deep learning techniques for ear diseases based on segmentation of the normal tympanic membrane. *Clin Exp Otorhinolaryngol* 2023;16(1):28-36 [FREE Full text] [doi: [10.21053/ceo.2022.00675](https://doi.org/10.21053/ceo.2022.00675)] [Medline: [36330706](https://pubmed.ncbi.nlm.nih.gov/36330706/)]
6. Alhudaif A, Cömert Z, Polat K. Otitis media detection using tympanic membrane images with a novel multi-class machine learning algorithm. *PeerJ Comput Sci* 2021;7:e405 [FREE Full text] [doi: [10.7717/peerj-cs.405](https://doi.org/10.7717/peerj-cs.405)] [Medline: [33817048](https://pubmed.ncbi.nlm.nih.gov/33817048/)]
7. Zeng X, Jiang Z, Luo W, Li H, Li H, Li G, et al. Efficient and accurate identification of ear diseases using an ensemble deep learning model. *Sci Rep* 2021;11(1):10839 [FREE Full text] [doi: [10.1038/s41598-021-90345-w](https://doi.org/10.1038/s41598-021-90345-w)] [Medline: [34035389](https://pubmed.ncbi.nlm.nih.gov/34035389/)]
8. Chen YC, Chu YC, Huang CY, Lee YT, Lee WY, Hsu CY, et al. Smartphone-based artificial intelligence using a transfer learning algorithm for the detection and diagnosis of middle ear diseases: a retrospective deep learning study. *EClinicalMedicine* 2022;51:101543 [FREE Full text] [doi: [10.1016/j.eclinm.2022.101543](https://doi.org/10.1016/j.eclinm.2022.101543)] [Medline: [35856040](https://pubmed.ncbi.nlm.nih.gov/35856040/)]
9. Brin D, Sorin V, Vaid A, Soroush A, Glicksberg BS, Charney AW, et al. Comparing ChatGPT and GPT-4 performance in USMLE soft skill assessments. *Sci Rep* 2023;13(1):16492 [FREE Full text] [doi: [10.1038/s41598-023-43436-9](https://doi.org/10.1038/s41598-023-43436-9)] [Medline: [37791711](https://pubmed.ncbi.nlm.nih.gov/37791711/)]
10. Jang D, Yun TR, Lee CY, Kwon YK, Kim CE. GPT-4 can pass the Korean national licensing examination for Korean medicine doctors. *PLOS Digit Health* 2023;2(12):e0000416 [FREE Full text] [doi: [10.1371/journal.pdig.0000416](https://doi.org/10.1371/journal.pdig.0000416)] [Medline: [38100393](https://pubmed.ncbi.nlm.nih.gov/38100393/)]
11. Takagi S, Watari T, Erabi A, Sakaguchi K. Performance of GPT-3.5 and GPT-4 on the Japanese medical licensing examination: comparison study. *JMIR Med Educ* 2023;9:e48002 [FREE Full text] [doi: [10.2196/48002](https://doi.org/10.2196/48002)] [Medline: [37384388](https://pubmed.ncbi.nlm.nih.gov/37384388/)]
12. GPT-4V(ision) system card. OpenAI. 2023. URL: <https://openai.com/research/gpt-4v-system-card> [accessed 2023-09-25]
13. Wu W, Yao H, Zhang M, Song Y, Ouyang W, Wang J. GPT4Vis: what can GPT-4 do for zero-shot visual recognition? arXiv Preprint posted online on March 12 2024. [doi: [10.48550/arXiv.2311.15732](https://doi.org/10.48550/arXiv.2311.15732)]
14. Chen RJ, Lu MY, Williamson DFK, Chen TY, Lipkova J, Noor Z, et al. Pan-cancer integrative histology-genomic analysis via multimodal deep learning. *Cancer Cell* 2022 08;40(8):865-878.e6 [FREE Full text] [doi: [10.1016/j.ccell.2022.07.004](https://doi.org/10.1016/j.ccell.2022.07.004)] [Medline: [35944502](https://pubmed.ncbi.nlm.nih.gov/35944502/)]
15. Zhang L, Jiang Y, Jin Z, Jiang W, Zhang B, Wang C, et al. Real-time automatic prediction of treatment response to transcatheter arterial chemoembolization in patients with hepatocellular carcinoma using deep learning based on digital

- subtraction angiography videos. *Cancer Imaging* 2022;22(1):23 [FREE Full text] [doi: [10.1186/s40644-022-00457-3](https://doi.org/10.1186/s40644-022-00457-3)] [Medline: [35549776](https://pubmed.ncbi.nlm.nih.gov/35549776/)]
16. Ming Y, Dong X, Zhao J, Chen Z, Wang H, Wu N. Deep learning-based multimodal image analysis for cervical cancer detection. *Methods* 2022;205:46-52 [FREE Full text] [doi: [10.1016/j.ymeth.2022.05.004](https://doi.org/10.1016/j.ymeth.2022.05.004)] [Medline: [35598831](https://pubmed.ncbi.nlm.nih.gov/35598831/)]
 17. Tanaka Y, Nakata T, Aiga K, Etani T, Muramatsu R, Katagiri S, et al. Performance of generative pretrained transformer on the national medical licensing examination in Japan. *PLoS Digit Health* 2024;3(1):e0000433 [FREE Full text] [doi: [10.1371/journal.pdig.0000433](https://doi.org/10.1371/journal.pdig.0000433)] [Medline: [38261580](https://pubmed.ncbi.nlm.nih.gov/38261580/)]
 18. Masao N, Takayoshi U, Ryota K, Mari SD, Ito M, Yamoto N, et al. A study of the performance of the generative pretrained transformer in the Japanese otorhinolaryngology specialty examination. *Nippon Jibiinkoka Tokeibugeka Gakkai Kaiho (Tokyo)* 2023;126:1217-1223 [FREE Full text] [doi: [10.3950/jibiinkotokeibu.126.11_1217](https://doi.org/10.3950/jibiinkotokeibu.126.11_1217)]
 19. Bharat SM, Myrzakhan A, Shen Z. Principled instructions are all you need for questioning LLaMA-1/2, GPT-3.5/4. *arXiv Preprint* posted online on January 18 2024. [doi: [10.48550/arXiv.2312.16171](https://doi.org/10.48550/arXiv.2312.16171)]
 20. Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, Agarwal S, et al. Learning transferable visual models from natural language supervision. *arXiv Preprint* posted online on February 26 2021. [doi: [10.48550/arXiv.2103.00020](https://doi.org/10.48550/arXiv.2103.00020)]
 21. Crowson MG, Bates DW, Suresh K, Cohen MS, Hartnick CJ. "Human vs Machine" validation of a deep learning algorithm for pediatric middle ear infection diagnosis. *Otolaryngol Head Neck Surg* 2023;169(1):41-46 [FREE Full text] [doi: [10.1177/01945998221119156](https://doi.org/10.1177/01945998221119156)] [Medline: [35972815](https://pubmed.ncbi.nlm.nih.gov/35972815/)]
 22. Suresh K, Wu MP, Benboujja F, Christakis B, Newton A, Hartnick CJ, et al. AI model versus clinician otoscopy in the operative setting for otitis media diagnosis. *Otolaryngol Head Neck Surg* 2023 (forthcoming) [FREE Full text] [doi: [10.1002/ohn.559](https://doi.org/10.1002/ohn.559)] [Medline: [37822130](https://pubmed.ncbi.nlm.nih.gov/37822130/)]
 23. Hu Y, Chen Q, Du J, Peng X, Keloth VK, Zuo X, et al. Improving large language models for clinical named entity recognition via prompt engineering. *J Am Med Inform Assoc* 2024:ocad259 (forthcoming) [FREE Full text] [doi: [10.1093/jamia/ocad259](https://doi.org/10.1093/jamia/ocad259)] [Medline: [38281112](https://pubmed.ncbi.nlm.nih.gov/38281112/)]
 24. Meskó B. Prompt engineering as an important emerging skill for medical professionals: tutorial. *J Med Internet Res* 2023;25:e50638 [FREE Full text] [doi: [10.2196/50638](https://doi.org/10.2196/50638)] [Medline: [37792434](https://pubmed.ncbi.nlm.nih.gov/37792434/)]
 25. Tseng CC, Lim V, Jyung RW. Use of artificial intelligence for the diagnosis of cholesteatoma. *Laryngoscope Investig Otolaryngol* 2023;8(1):201-211 [FREE Full text] [doi: [10.1002/liv.1008](https://doi.org/10.1002/liv.1008)] [Medline: [36846416](https://pubmed.ncbi.nlm.nih.gov/36846416/)]
 26. Kanjee Z, Crowe B, Rodman A. Accuracy of a generative artificial intelligence model in a complex diagnostic challenge. *JAMA* 2023;330(1):78-80 [FREE Full text] [doi: [10.1001/jama.2023.8288](https://doi.org/10.1001/jama.2023.8288)] [Medline: [37318797](https://pubmed.ncbi.nlm.nih.gov/37318797/)]
 27. Truhn D, Weber CD, Braun BJ, Bressemer K, Kather JN, Kuhl C, et al. A pilot study on the efficacy of GPT-4 in providing orthopedic treatment recommendations from MRI reports. *Sci Rep* 2023;13(1):20159 [FREE Full text] [doi: [10.1038/s41598-023-47500-2](https://doi.org/10.1038/s41598-023-47500-2)] [Medline: [37978240](https://pubmed.ncbi.nlm.nih.gov/37978240/)]
 28. Vlahou A, Hallinan D, Apweiler R, Argiles A, Beige J, Benigni A, et al. Data sharing under the general data protection regulation: time to harmonize law and research ethics? *Hypertension* 2021;77(4):1029-1035 [FREE Full text] [doi: [10.1161/HYPERTENSIONAHA.120.16340](https://doi.org/10.1161/HYPERTENSIONAHA.120.16340)] [Medline: [33583200](https://pubmed.ncbi.nlm.nih.gov/33583200/)]
 29. Fournier-Tombs E, McHardy J. A medical ethics framework for conversational artificial intelligence. *J Med Internet Res* 2023;25:e43068 [FREE Full text] [doi: [10.2196/43068](https://doi.org/10.2196/43068)] [Medline: [37224277](https://pubmed.ncbi.nlm.nih.gov/37224277/)]
 30. Editage. URL: <https://www.editage.com/> [accessed 2024-05-11]

Abbreviations

AI: artificial intelligence
AOM: acute otitis media
chole: middle ear cholesteatoma
COM: chronic otitis media
FN: false negative
FP: false positive
GPT-4V: GPT-4 Vision
LLM: large language model
OME: otitis media with effusion
TP: true positive

Edited by Y Huo; submitted 13.03.24; peer-reviewed by S Murono, B Li, J Jagtap; comments to author 10.04.24; revised version received 23.04.24; accepted 07.05.24; published 31.05.24.

Please cite as:

Noda M, Yoshimura H, Okubo T, Koshu R, Uchiyama Y, Nomura A, Ito M, Takumi Y

Feasibility of Multimodal Artificial Intelligence Using GPT-4 Vision for the Classification of Middle Ear Disease: Qualitative Study and Validation

JMIR AI 2024;3:e58342

URL: <https://ai.jmir.org/2024/1/e58342>

doi: [10.2196/58342](https://doi.org/10.2196/58342)

PMID: [38875669](https://pubmed.ncbi.nlm.nih.gov/38875669/)

©Masao Noda, Hidekane Yoshimura, Takuya Okubo, Ryota Koshu, Yuki Uchiyama, Akihiro Nomura, Makoto Ito, Yutaka Takumi. Originally published in JMIR AI (<https://ai.jmir.org>), 31.05.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR AI, is properly cited. The complete bibliographic information, a link to the original publication on <https://www.ai.jmir.org/>, as well as this copyright and license information must be included.

Original Paper

Evaluating the Diagnostic Performance of Symptom Checkers: Clinical Vignette Study

Mohammad Hammoud¹, PhD; Shahd Douglas¹, MSc; Mohamad Darmach¹, MD; Sara Alawneh¹, MD; Swapnendu Sanyal¹, MSc; Youssef Kanbour¹, BSc

Avey Inc, Doha, Qatar

Corresponding Author:

Mohammad Hammoud, PhD

Avey Inc

Qatar Science and Technology Park

Doha, 210022

Qatar

Phone: 974 3001 8035

Email: mhh@avey.ai

Abstract

Background: Medical self-diagnostic tools (or symptom checkers) are becoming an integral part of digital health and our daily lives, whereby patients are increasingly using them to identify the underlying causes of their symptoms. As such, it is essential to rigorously investigate and comprehensively report the diagnostic performance of symptom checkers using standard clinical and scientific approaches.

Objective: This study aims to evaluate and report the accuracies of a few known and new symptom checkers using a standard and transparent methodology, which allows the scientific community to cross-validate and reproduce the reported results, a step much needed in health informatics.

Methods: We propose a 4-stage experimentation methodology that capitalizes on the standard clinical vignette approach to evaluate 6 symptom checkers. To this end, we developed and peer-reviewed 400 vignettes, each approved by at least 5 out of 7 independent and experienced primary care physicians. To establish a frame of reference and interpret the results of symptom checkers accordingly, we further compared the best-performing symptom checker against 3 primary care physicians with an average experience of 16.6 (SD 9.42) years. To measure accuracy, we used 7 standard metrics, including M1 as a measure of a symptom checker's or a physician's ability to return a vignette's main diagnosis at the top of their differential list, F_1 -score as a trade-off measure between recall and precision, and Normalized Discounted Cumulative Gain (NDCG) as a measure of a differential list's ranking quality, among others.

Results: The diagnostic accuracies of the 6 tested symptom checkers vary significantly. For instance, the differences in the M1, F_1 -score, and NDCG results between the best-performing and worst-performing symptom checkers or ranges were 65.3%, 39.2%, and 74.2%, respectively. The same was observed among the participating human physicians, whereby the M1, F_1 -score, and NDCG ranges were 22.8%, 15.3%, and 21.3%, respectively. When compared against each other, physicians outperformed the best-performing symptom checker by an average of 1.2% using F_1 -score, whereas the best-performing symptom checker outperformed physicians by averages of 10.2% and 25.1% using M1 and NDCG, respectively.

Conclusions: The performance variation between symptom checkers is substantial, suggesting that symptom checkers cannot be treated as a single entity. On a different note, the best-performing symptom checker was an artificial intelligence (AI)-based one, shedding light on the promise of AI in improving the diagnostic capabilities of symptom checkers, especially as AI keeps advancing exponentially.

(JMIR AI 2024;3:e46875) doi:[10.2196/46875](https://doi.org/10.2196/46875)

KEYWORDS

digital health; symptom checker; artificial intelligence; AI; patient-centered care; eHealth apps; eHealth

Introduction

Background

Digital health has become ubiquitous. Every day, millions of people turn to the internet for health information and treatment advice [1,2]. For instance, in Australia, approximately 80% of people search the internet for health information and approximately 40% seek web-based guidance for self-treatment [3,4]. In the United States, approximately two-thirds of adults search the web for health information and one-third use it for self-diagnosis, trying to singlehandedly understand the underlying causes of their health symptoms [5]. A recent study showed that half of the patients investigated their symptoms on search engines before visiting emergency rooms [6,7].

Although search engines such as Google and Bing are exceptional tools for educating people on almost any matter, they may facilitate misdiagnosis and induce serious risks [5]. This is because searching the web entails sifting through a great deal of information, stemming from all kinds of sources, and making personal medical judgments, correlations, and deductions accordingly. Some governments have even launched “Don’t Google It” advertising campaigns to raise public awareness of the risks of assessing one’s health using search engines [8,9]. The reality is that search engines are not medical diagnostic tools and laymen are not usually equipped to leverage them for self-diagnosis.

In contrast to search engines, symptom checkers are patient-facing medical diagnostic tools that emulate clinical reasoning, especially if they use artificial intelligence (AI) [4,10]. They are trained to make medical expert-like judgments on behalf of patients. More precisely, a patient can start a consultation session with a symptom checker by inputting a chief complaint (in terms of ≥ 1 symptoms). Afterward, the symptom checker asks several questions to the patient and collects answers from them. Finally, it generates a differential diagnosis (ie, a ranked list of potential diseases) that explains the causes of the patient’s symptoms.

Symptom checkers are increasingly becoming an integral part of digital health, with >15 million people using them on a monthly basis [11], a number that is expected to continue to grow [12]. A United Kingdom-based study [13] that engaged 1071 patients found that >70% of individuals aged between 18 and 39 years would use a symptom checker. A recent study examining a specific symptom checker found that >80% of patients perceived it to be useful and >90% indicated that they would use it again [14]. Various credible health care institutions and entities such as the UK National Health Service [15] and the government of Australia [16] have officially adopted symptom checkers for self-diagnosis and referrals.

Symptom checkers are inherently scalable (ie, they can assess millions of people instantly and concurrently) and universally available. In addition, they promise to provide patients with necessary high-quality, evidence-based information [17]; reduce unnecessary medical visits [18-21]; alleviate the pressure on health care systems [22]; improve accessibility to timely

diagnosis [18]; and guide patients to the most appropriate care pathways [12], to mention just a few.

Nevertheless, the utility and promise of symptom checkers cannot be materialized if they are not proven to be accurate [10]. To elaborate, a recent study has shown that most patients (>76%) use symptom checkers solely for self-diagnosis [14]. As such, if symptom checkers are not meticulously engineered and rigorously evaluated on their diagnostic capabilities, they may put patients at risk [23-25].

This study investigates the diagnostic performance of symptom checkers by measuring the accuracies of a few popular symptom checkers and a new AI-based symptom checker. In addition, it compares the accuracy of the best-performing symptom checker against that of a panel of experienced physicians to put things in perspective and interpret results accordingly.

Evaluation Methodology

To evaluate symptom checkers, we propose a scientific methodology that capitalizes on the standard clinical vignette approach [26] (Multimedia Appendix 1 provides additional information on how our methodology aligns with the recommended requirements of this approach [4,7,12,26-39]). Delivering on this methodology, we compiled 400 vignettes and peer reviewed them with 7 external physicians using a supermajority voting scheme. To the best of our knowledge, this yielded the largest benchmark vignette suite in the domain thus far. Furthermore, we defined and used 7 standard accuracy metrics, one of which measures for the first time, the ranking qualities of the differential diagnoses of symptom checkers and physicians.

Subsequently, we leveraged the peer-reviewed benchmark vignette suite and accuracy metrics to investigate the performance of a new AI-based symptom checker named Avey [40] and 5 popular symptom checkers named Ada [41], K Health [42], Buoy [43], Babylon [44], and WebMD [45]. Results demonstrated a significant performance variation between these symptom checkers and the promise of AI in improving their diagnostic capabilities. For example, the best-performing symptom checker, namely Avey, outperformed Ada, K Health, Buoy, Babylon, and WebMD by averages of 24.5%, 142.8%, 159.6%, 2968.1%, and 175.5%, respectively, in listing the vignettes’ main diagnoses at the top of their differentials.

Avey claims to use advanced AI technology [40]. In particular, it involves a diagnostic engine that operationalizes a probabilistic graphical model, namely a Bayesian network. Figure 1 demonstrates the model in action, which was built bottom-up over 4 years specifically for medical diagnosis. In addition, the engine uses a recommendation system, which predicts the future impact of every symptom or etiology that has not yet been asked during a patient session with Avey and recommends the one that exhibits the highest impact on the engine’s current diagnostic hypothesis. At the end of the session, a ranking model is used for ranking all the possible diseases for the patient’s case and outputs them as a differential diagnosis.

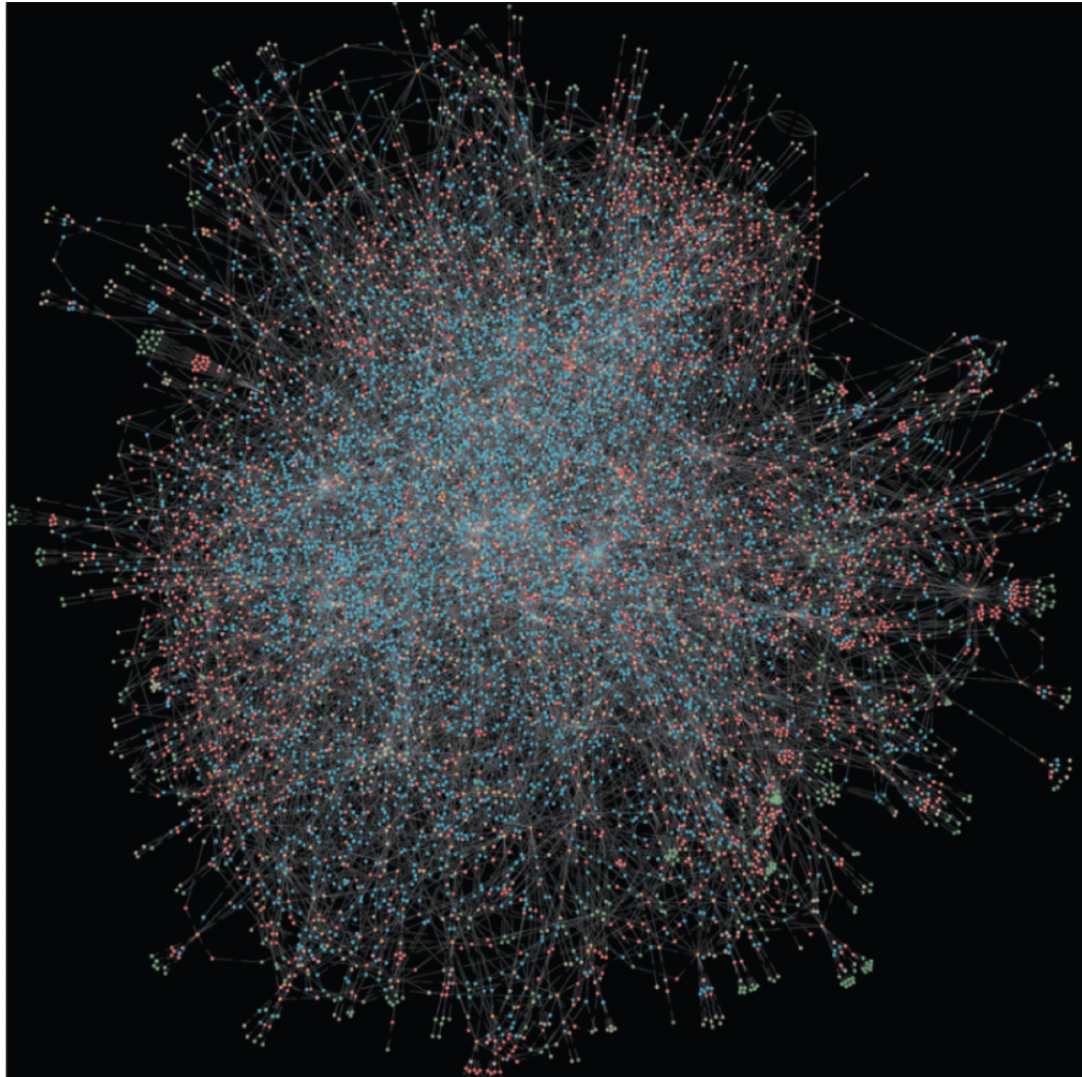
To put things in perspective, we subsequently compared the performance of Avey against 3 primary care physicians with an average experience of 16.6 years. The results showed that

Avey compared favorably to the physicians and slightly outperformed them in some accuracy metrics, including the ability to rank diseases correctly within their generated differential lists.

Finally, to facilitate the reproducibility of the study and support future related studies, we made the peer-reviewed benchmark

vignette suite publicly and freely available [27]. In addition, we posted all the results of the symptom checkers and physicians in the Benchmark Vignette Suite [27] to establish a standard of full transparency and allow researchers to cross-validate the results, a step much needed in health informatics [46].

Figure 1. An actual visualization of Avey's brain (ie, a probabilistic graphical model). At a high level, the nodes (or dots) can be thought of representing diseases, symptoms, etiologies, or features of symptoms or etiologies, whereas the edges (or links) can be thought of as representing conditional independence assumptions and modeling certain features (eg, sensitivities and specificities) needed for clinical reasoning.



Methods

Stages

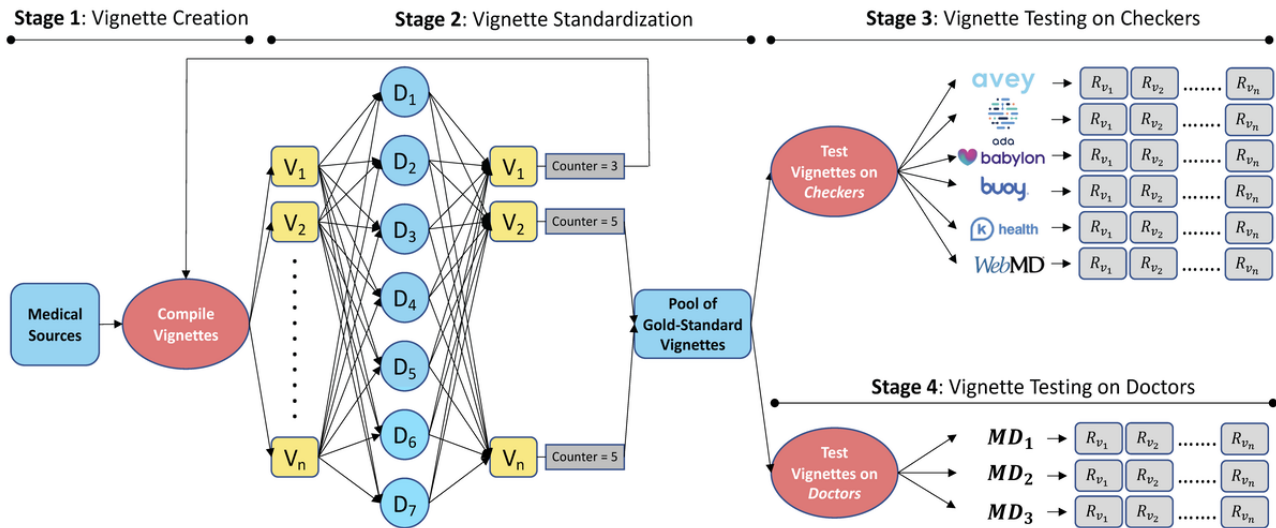
Overview

Building on prior related work [4,5,11,12,26,28,29], we adopted a clinical vignette approach to measure the performance of symptom checkers. A seminal work at Harvard Medical School has established the value of this approach in validating the

accuracies of symptom checkers [11,29], especially because it has been also used as a common approach to test physicians on their diagnostic capabilities [29].

To this end, we defined our experimentation methodology in terms of 4 stages, namely *vignette creation*, *vignette standardization*, *vignette testing on symptom checkers*, and *vignette testing on doctors*. The 4 stages are illustrated in [Figure 2](#).

Figure 2. Our 4-stage experimentation methodology (V_i =vignette i , assuming n vignettes and $1 \leq i \leq n$; D_j =doctor j , assuming 7 doctors and $1 \leq j \leq 7$; MD_k =medical doctor k , assuming 3 doctors and $1 \leq k \leq 3$; R_i =result of vignette V_i as generated by a checker or a medical doctor [MD]). In the “vignette creation” stage, the vignettes are compiled from reputable medical sources by an internal team of MDs. In the “vignette standardization” stage, the vignettes are reviewed and approved by a panel of experienced and independent physicians. In the “vignette testing on symptom checkers” stage, the vignettes are tested on symptom checkers by a different panel of experienced and independent physicians. In the “vignette testing on doctors” stage, the vignettes are tested on a yet different panel of experienced and independent physicians.



Stage 1: Vignette Creation Stage

In this stage, an internal team of 3 physicians (akin to the study by Gilbert et al [28]) compiled a set of vignettes from October 10, 2021, to November 29, 2021. All the vignettes were drawn from reputable medical websites and training material for health care professionals, including the United States Medical Licensing Examination, Step 2 CK, Membership of the Royal Colleges of Physicians Part 1 Self-Assessment, American Board of Family Medicine, and American Board of Pediatrics, among others [30-37]. In addition, the internal medical team supplemented the vignettes with information that might be “asked” by symptom checkers and physicians in stages 3 and 4. The vignettes involved 14 body systems and encompassed common and less-common conditions relevant to primary care

practice (Table 1). They fairly represent real-life or practical cases in which patients might seek primary care advice from physicians or symptom checkers.

The internal medical team constructed each vignette in terms of eight major components: (1) the age and sex of the assumed patient; (2) a maximum of 3 chief complaints; (3) the history of the suggested illness associated with details on the chief complaints and other present and relevant findings (a finding is defined as a symptom, a sign, or an etiology, each with a potential attribute); (4) absent findings, including ones that are expected to be solicited by symptom checkers and physicians in stages 3 and 4; (5) basic findings that pertain to physical examinations that can still be exploited by symptom checkers; (6) past medical and surgical history; (7) family history; and (8) the most appropriate main and differential diagnoses.

Table 1. The body systems and numbers of common and less-common diseases covered in the compiled vignette suite.

Body system	Vignettes			Covered diseases, % (p ^a /P ^b)
	Weightage in the suite, % (n ^c /N ^d)	Vignettes with common diseases % (m ^e /n) (total: 55.5%, 222/400)	Vignettes with less-common diseases, % (k ^f /n) (total: 44.5%, 178/400)	
Hematology	5.75 (23/400)	8.7 (2/23)	91.3 (21/23)	4.89 (13/266)
Cardiovascular	11.5 (46/400)	58.7 (27/46)	41.3 (19/46)	11.28 (30/266)
Neurology	5.5 (22/400)	40.91 (9/22)	59.09 (13/22)	5.26 (14/266)
Endocrine	20 (5/5) (20/400)	65 (13/20)	35 (7/20)	4.89 (13/266)
ENT ^g	5.75 (23/400)	69.57 (16/23)	30.43 (7/23)	5.64 (15/266)
GI ^h	11 (44/400)	47.73 (21/44)	52.27 (23/44)	12.78 (34/266)
Obstetrics and gynecology	13.5 (54/400)	59.26 (32/54)	40.74 (22/54)	13.16 (35/266)
Infectious	5.75 (23/400)	26.09(6/23)	73.91 (17/23)	6.39 (17/266)
Respiratory	9.25 (37/400)	70.27 (26/37)	29.73 (11/37)	7.52 (20/266)
Orthopedics and rheumatology	8 (32/400)	65.63 (21/32)	34.38 (11/32)	9.4 (25/266)
Ophthalmology	4.5 (18/400)	83.33 (15/18)	16.67 (3/18)	4.51 (12/266)
Dermatology	3 (12/400)	75 (9/12)	25 (3/12)	4.51 (12/266)
Urology	3.5 (14/400)	57.14 (8/14)	42.86 (6/14)	3.01 (8/266)
Nephrology	8 (32/400)	53.13 (17/32)	46.88 (15/32)	6.77 (18/266)

^ap: number of diseases covered in the body system.

^bP: total number of diseases covered by the N vignettes.

^cn: number of vignettes for the corresponding body system.

^dN: total number of vignettes in our suite.

^em: count of vignettes covering common diseases of the corresponding body system.

^fk: count of vignettes covering less-common diseases of the corresponding body system.

^gENT: ear, nose, and throat.

^hGI: gastrointestinal.

Stage 2: Vignette Standardization Stage

The output of the vignette creation stage (ie, stage 1) is a set of vignettes that serves as an input to the vignette standardization stage (ie, stage 2). Seven external physicians (as opposed to 3 doctors in the study by Gilbert et al [28]) from 4 specialties, namely family medicine, general medicine, emergency medicine, and internal medicine, with an average experience of 8.4 years were recruited from the professional networks of the authors to review the vignettes in this stage. None of these external doctors had any involvement with the development of any of the symptom checkers considered in this study.

We designed and developed a full-fledged web portal to streamline the process of reviewing and standardizing the vignettes. To elaborate, the portal allows the internal medical team to upload the vignettes to a web page that is shared across the 7 externally recruited doctors. Each doctor can access the vignettes and review them independently, without seeing the reviews of other doctors.

After reviewing a vignette, a doctor can reject or accept it. Upon rejecting a vignette, a doctor can propose changes to improve its quality or clarity. The internal medical team checks the

suggested changes, updates the vignette accordingly, and reuploads it to the portal for a new round of peer reviewing by the 7 external doctors. Multiple reviewing rounds can take place before a vignette is rendered gold standard. A vignette becomes the gold standard only if it is accepted by at least 5 out of the 7 (ie, supermajority) external doctors. Once a vignette is standardized, the portal moves it automatically to stages 3 and 4.

Stage 2 started on October 17, 2021, and ended on December 4, 2021. As an outcome, 400 vignettes were produced and standardized. To allow for external validation, we made all the vignettes publicly available [27].

Stage 3: Vignette Testing on Symptom Checkers

The output of stage 2 serves as an input to stage 3, namely, vignette testing on symptom checkers. For this sake, we recruited 3 independent primary care physicians under 2 specialties, namely family medicine and general medicine, with an average experience of 4.2 years from the professional networks of the authors. None of these physicians had any involvement with the development of any of the symptom checkers tested in this study. Furthermore, 2 of them were not among the 7 doctors who reviewed the vignettes in stage 2.

These doctors were recruited solely to test the gold-standard vignettes on the considered symptom checkers.

The approach of having primary care physicians test symptom checkers has been shown recently to be more reliable than having laypeople do so [28,38,47]. This is because the standardized vignettes act as proxies for patients, whereas testers act as only data extractors from the vignettes and information feeders to the symptom checkers. Consequently, the better the testers are in extracting and feeding data, the more reliable the clinical vignette approach renders. In fact, a symptom checker cannot be judged on its accuracy if the answers to its questions are not in full alignment with the contents of the vignettes.

To this end, physicians are deemed more capable of playing the role of testers than laypeople, especially that AI-based symptom checkers (eg, Ada and Avey, among others) may often ask questions that have no answers in the vignettes, even if the vignettes are quite comprehensive. Clearly, when these questions are asked, laypeople will not be able to answer them properly, impacting thereby the reliability of the clinical vignette approach and the significance of the reported results. In contrast, physicians will judiciously answer these questions in alignment with the vignettes and capably figure out whether the symptom checkers are able to “diagnose” them (ie, produce the correct differential diagnoses in the vignettes). We elaborate further on the rationale behind using physicians as testers in the Strengths and Limitations section.

Besides vignettes, we chose 6 symptom checkers, namely Ada [41], Babylon [44], Buoy [43], K Health [42], WebMD [45], and Avey [40], to evaluate their performance and compare them against each other. Four of these symptom checkers (ie, Ada, Buoy, K Health, and WebMD) were selected because of their superior performance reported in Gilbert et al [28], and 1 (ie, Babylon) was chosen because of its popularity. Avey is a new AI-based symptom checker that is emerging, with >1 million people who have already downloaded it [40]. We tested the gold-standard vignettes on the most up-to-date versions of these symptom checkers that were available on Google Play, App Store, or websites (eg, Buoy) between the dates of November 7, 2021, and January 31, 2022.

The 6 symptom checkers were tested through their normal question-answer flows. As in the study by Gilbert et al [28], each of the external physicians in stage 3 randomly pulled vignettes from the gold-standard pool and tested them on *each* of the 6 symptom checkers (compared to the study by Gilbert et al [28], where 8 doctors tested vignettes on 4 symptom checkers; Figure 2). By the end of stage 3, each physician tested a total of 133 gold-standard vignettes on each symptom checker, except 1 physician who tested 1 extra vignette to exhaust the 400 vignettes. Each physician saved a screenshot of each symptom checker’s output for each vignette to facilitate the results’ verification, extraction, and analysis. We posted all the screenshots on the internet [27] to establish a standard of full transparency and allow for external cross-validation and study replication.

Stage 4: Vignette Testing on Doctors

In this stage, we recruited 3 more independent and experienced primary care physicians with an average experience of 16.6 years (compared with 7 doctors in the study by Gilbert et al [28], with an average experience of 11.2 years) from the professional networks of the authors. One of those physicians is a family medicine doctor with >30 years of experience. The other 2 are also family medicine doctors, each with >10 years of experience. None of these physicians had any involvement with the development of any of the tested symptom checkers. Furthermore, none of them was among the 7 or 3 doctors of stages 2 or 3, respectively, and they were all only recruited to pursue stage 4.

The sole aim of stage 4 is to compare the accuracy of the winning symptom checker against that of experienced primary care physicians. Hence, similar to the study by Semigran et al [11], we concealed the main and differential diagnoses of the 400 gold-standard vignettes from the 3 recruited doctors and exposed the remaining information through our web portal. The doctors were granted access to the portal and asked to provide their main and differential diagnoses for each vignette without checking any reference, mimicking as closely as possible the way they conduct real-world sessions live with patients. As an outcome, each vignette was “diagnosed” by each of the 3 doctors. The results of the doctors were posted to allow for external cross-validation [27].

Finally, we note that different symptom checkers and doctors can refer to the same disease differently. As such, we considered an output disease by a symptom checker (in stage 3) or a doctor (in stage 4) as a reasonable match to a disease in the gold-standard vignette if it was an alternative name, an umbrella name, or a directly related disease.

Accuracy Metrics

To evaluate the performance of symptom checkers and doctors in stages 3 and 4, we used 7 standard accuracy metrics. As in the study by Gilbert et al [28] and United States Medical Licensing Examination [48], for every tested gold-standard vignette, we used the matching-1 ($M1$), matching-3 ($M3$), and matching-5 ($M5$) criteria to measure if a symptom checker or a doctor is able to output the vignette’s main diagnosis at the top (ie, $M1$), among the first 3 diseases (ie, $M3$), or among the first 5 diseases (ie, $M5$) of their differential list. For each symptom checker and doctor, we report the percentages of vignettes that fulfill $M1$, $M3$, and $M5$. The mathematical definitions of $M1$, $M3$, and $M5$ are given in Table 2.

Besides, as in the studies by Gilbert et al [28], Baker et al [38], and Kannan et al [49], for each tested gold-standard vignette, we used recall (or sensitivity in medical parlance) as a measure of the percentage of relevant diseases that are returned in the symptom checker’s or doctor’s differential list. Moreover, we used precision as a measure of the percentage of diseases in the symptom checker’s or doctor’s differential list that are relevant. For each symptom checker and doctor, we report the average recall and average precision (see Table 2 for their mathematical definitions) across all vignettes.

Typically, there is a trade-off between recall and precision (the higher the recall, the lower the precision, and vice versa). Thus, in accordance with the standard practice in computer science, we further used the F_1 -measure that combines the trade-off

between recall and precision in one easily interpretable score. The mathematical definition of the F_1 -measure is provided in Table 2. The higher the F_1 -measure of a symptom checker or a doctor, the better.

Table 2. The descriptions and mathematical definitions of the 7 accuracy metrics used in this study.

Metric	Description	Mathematical definition
M1%	The percentage of vignettes where the gold standard main diagnosis is returned at the top of a symptom checker's or a doctor's differential list	$\frac{1}{N} \sum_{v=1}^N i_v$, where N is the number of vignettes and i_v is 1 if the symptom checker or doctor returns the gold standard main diagnosis within vignette v at the top of their differential list; and 0 otherwise
M3%	The percentage of vignettes where the gold standard main diagnosis is returned among the first 3 diseases of a symptom checker's or a doctor's differential list	$\frac{1}{N} \sum_{v=1}^N i_v$, where N is the number of vignettes and i_v is 1 if the symptom checker or doctor returns the gold standard main diagnosis within vignette v among the top 3 diseases of their differential list; and 0 otherwise
M5%	The percentage of vignettes where the gold standard main diagnosis is returned among the first 5 diseases of a symptom checker's or a doctor's differential list	$\frac{1}{N} \sum_{v=1}^N i_v$, where N is the number of vignettes and i_v is 1 if the symptom checker or doctor returns the gold standard main diagnosis within vignette v among the top 5 diseases of their differential list; and 0 otherwise
Average recall	Recall is the proportion of diseases that are in the gold standard differential list and are generated by a symptom checker or a doctor. The average recall is taken across all vignettes for each symptom checker and doctor	$\frac{1}{N} \sum_{v=1}^N r_v$, where N is the number of vignettes and r_v of the symptom checker or doctor for vignette v
Average precision	Precision is the proportion of diseases in the symptom checker's or doctor's differential list that are also in the gold standard differential list. The average precision is taken across all vignettes for each symptom checker and doctor	$\frac{1}{N} \sum_{v=1}^N p_v$, where N is the number of vignettes and p_v of the symptom checker or doctor for vignette v
Average F_1 -measure	F_1 -measure captures the trade-off between precision and recall. The average F_1 -measure is taken across all vignettes for each symptom checker and doctor	$\frac{1}{N} \sum_{v=1}^N F_1$, where <i>average recall</i> and <i>average precision</i> are as defined at column 3 in rows 4 and 5 above, respectively
Average NDCG ^a	NDCG is a measure of ranking quality. The average NDCG is taken across all vignettes for each symptom checker and doctor	$\frac{1}{N} \sum_{v=1}^N \text{NDCG}_v$, assuming N vignettes, n number of diseases in a gold standard vignette v , and r_j relevance _{j} for the disease at position j in v 's differential list DCG_v , which is computed over the differential list of a doctor or a symptom checker for v . <i>Gold DCG_v</i> is defined exactly as DCG_v , but is computed over the gold standard differential list of v

^aNDCG: Normalized Discounted Cumulative Gain.

Finally, we measured the ranking qualities of each symptom checker and doctor using the Normalized Discounted Cumulative Gain (NDCG) [50] metric that is widely used in practice [51]. To begin with, each disease at position in the differential list of a gold-standard vignette is assigned . The higher the rank of a disease in the differential list, the higher the relevance of that disease to the correct diagnosis (eg, if a gold-standard differential has 2 diseases D1 and D2 in this order, they will be assigned relevancies 2 and 1, respectively). Next, Discounted Cumulative Gain (DCG) is defined mathematically as $\sum_{j=1}^n \frac{r_j}{\sqrt{j}}$, assuming diseases in a vignette's differential list (Table 2). As such, DCG penalizes a symptom checker or a doctor if they rank a disease lower in their output differential list than the gold-standard list. Capitalizing on DCG, NDCG is the ratio of a symptom checker's or a doctor's DCG divided by the

corresponding gold-standard DCG. Table 2 provides the mathematical definition of NDCG.

Ethical Considerations

No patients (whether as *subjects* or *testers*) were involved in any part of this study, but rather vignettes that acted as proxies for patients during testing with symptom checkers and physicians. As such, the vignettes are the subjects in this study and not humans. In addition, doctors were not subjects in stage 4 of the study (or any stage as a matter of fact), but rather the vignettes themselves. When the subjects are not humans, no institutional review board approval is typically required as per the guidelines of the United States Office for Human Research Protections [52]. This closely aligns with many of the related studies that use the clinical vignette approach [12,28,29,38,53,54], whereby none of them (to the best of our

knowledge) has obtained an institutional review board approval to conduct the study.

Results

Accuracies of Symptom Checkers

In this section, we present our findings of stage 3. As indicated in the Methods section, the 400 gold-standard vignettes were tested over 6 symptom checkers, namely Avey, Ada, WebMD, K Health, Buoy, and Babylon. Not every vignette was successfully diagnosed by every symptom checker. For instance, 18 vignettes failed on K Health because their constituent chief complaints were not available in K Health's search engine; hence, the sessions could not be initiated. Moreover, 35 vignettes failed on K Health because of an age limitation (only vignettes that encompassed ages of ≥ 18 years were accepted by K Health).

In addition to search and age limitations, some symptom checkers (in particular, Buoy) crashed while diagnosing certain vignettes, even after trying multiple times. Moreover, many symptom checkers did not produce differential diagnoses for some vignettes albeit concluding the diagnostic sessions. For example, Babylon did not generate differential diagnoses for 351 vignettes. The reason some symptom checkers could not produce diagnoses for some vignettes is uncertain, but we conjecture that it might relate to either not modeling those diagnoses or falling short of recalling them despite being modeled. Table 3 summarizes the failure rates and reasons across the examined symptom checkers. Moreover, the table shows the average number of questions asked by each symptom checker upon successfully diagnosing vignettes.

Table 3. Failure reasons, failure counts, success counts, and average number of questions across the 6 tested symptom checkers.

Symptom checker	Failure reasons and counts			Success counts		Number of questions, mean (SD)
	Search limitations	Age limitations	Crashed	No DDx ^a generated	DDx generated	
Avey	0	0	0	2	398	24.89 (12.15)
Ada	0	0	0	0	400	29.33 (6.62)
WebMD	2	1	0	3	394	2.64 (2.11)
K Health	18	35	0	2	345	25.23 (6.59)
Buoy	2	3	5	74	316	25.67 (5.79)
Babylon	15	0	0	351	34	5.91 (5.47)

^aDDx: differential diagnosis.

Figure 3 demonstrates the accuracy results of all the symptom checkers over the 400 vignettes, irrespective of whether they failed or not during some diagnostic sessions. In this set of results, a symptom checker is penalized if it fails to start a session, crashes, or does not produce a differential diagnosis albeit concluding the session. As depicted, Avey outperformed Ada, WebMD, K Health, Buoy, and Babylon, respectively, by averages of 24.5%, 175.5%, 142.8%, 159.6%, and 2968.1% using *M1*; 22.4%, 114.5%, 123.8%, 118.2%, and 3392% using *M3*; 18.1%, 79.2%, 116.8%, 125%, and 3114.2% using *M5*; 25.2%, 65.6%, 109.4%, 154%, and 3545% using recall; 8.7%, 88.9%, 66.4%, 88.9%, and 2084% using F_1 -measure; and 21.2%, 93.4%, 113.3%, 136.4%, and 3091.6% using NDCG. Ada was able to surpass Avey by an average of 0.9% using precision, although Avey outpaced it across all the remaining metrics, even with asking an average of 17.2% lesser number of questions (Table 3). As shown in Figure 3, Avey also outperformed WebMD, K Health, Buoy, and Babylon by

averages of 103.2%, 40.9%, 49.6%, and 1148.5% using precision, respectively.

Figure 4 illustrates the accuracy results of all the symptom checkers across only the vignettes that were successful. In other words, symptom checkers were not penalized if they failed to start sessions or crashed during sessions. As shown in the figure, Avey outperformed Ada, WebMD, K Health, Buoy, and Babylon, respectively, by averages of 24.5%, 173.2%, 110.9%, 152.8%, and 2834.7% using *M1*; 22.4%, 112.4%, 94%, 112.9%, and 3257.6% using *M3*; 18.1%, 77.8%, 88.2%, 119.5%, and 3003.4% using *M5*; 25.2%, 64.5%, 81.8%, 147.1%, and 3371.4% using recall; 8.7%, 87.6%, 44.4%, 83.8%, and 1922.2% using F_1 -measure; and 21.2%, 91.9%, 85%, 130.7%, and 2964% using NDCG. Under average precision, Ada outpaced Avey by an average of 0.9%, whereas Avey surpassed WebMD, K Health, Buoy, and Babylon by averages of 101.3%, 22%, 45.6%, and 1113.8%, respectively.

Figure 3. Accuracy results considering for each symptom checker all the succeeded and failed vignettes. NDCG: Normalized Discounted Cumulative Gain.

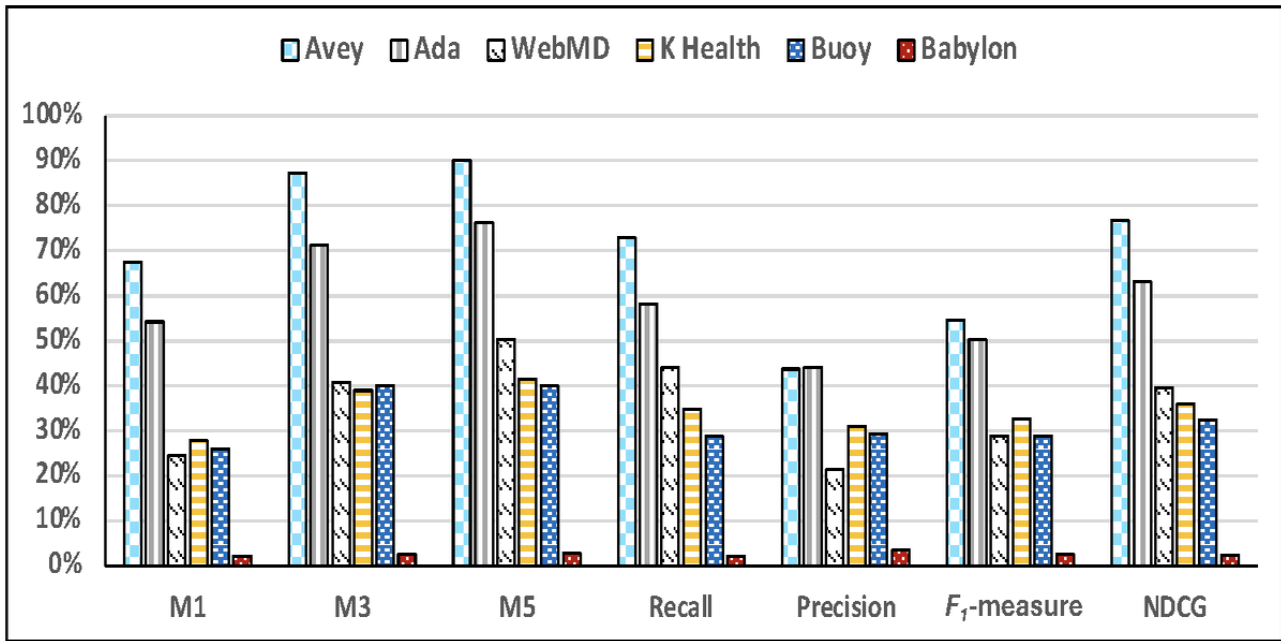
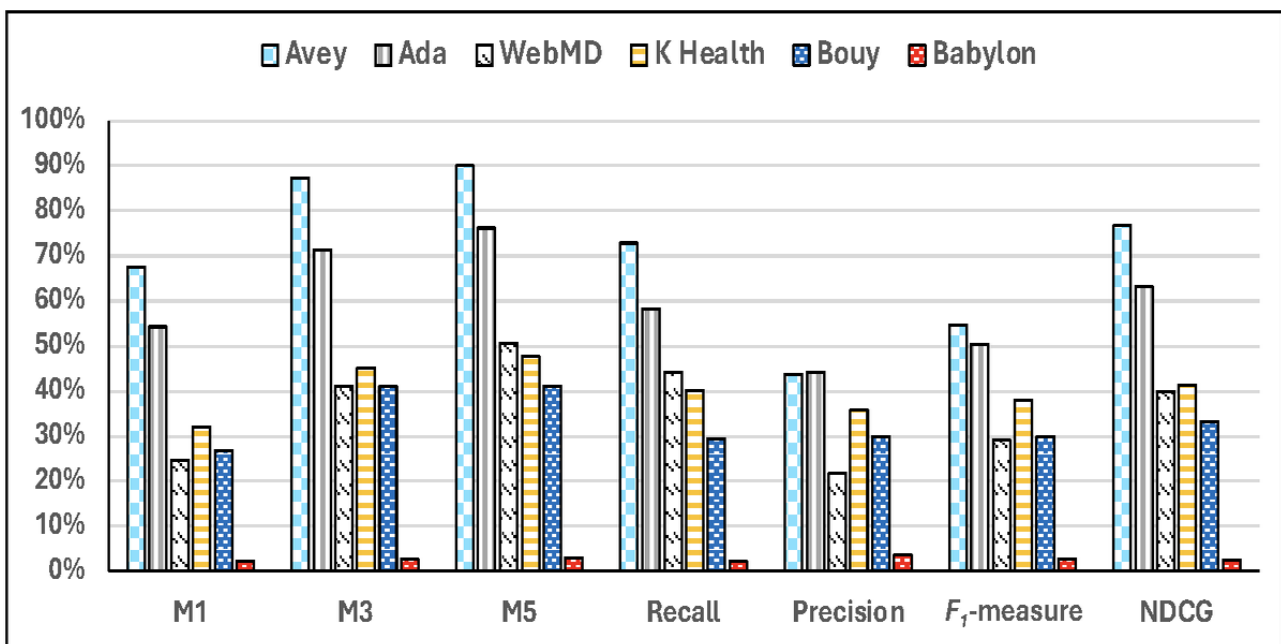


Figure 4. Accuracy results considering for each symptom checker only the succeeded vignettes, with or without differential diagnoses. NDCG: Normalized Discounted Cumulative Gain.

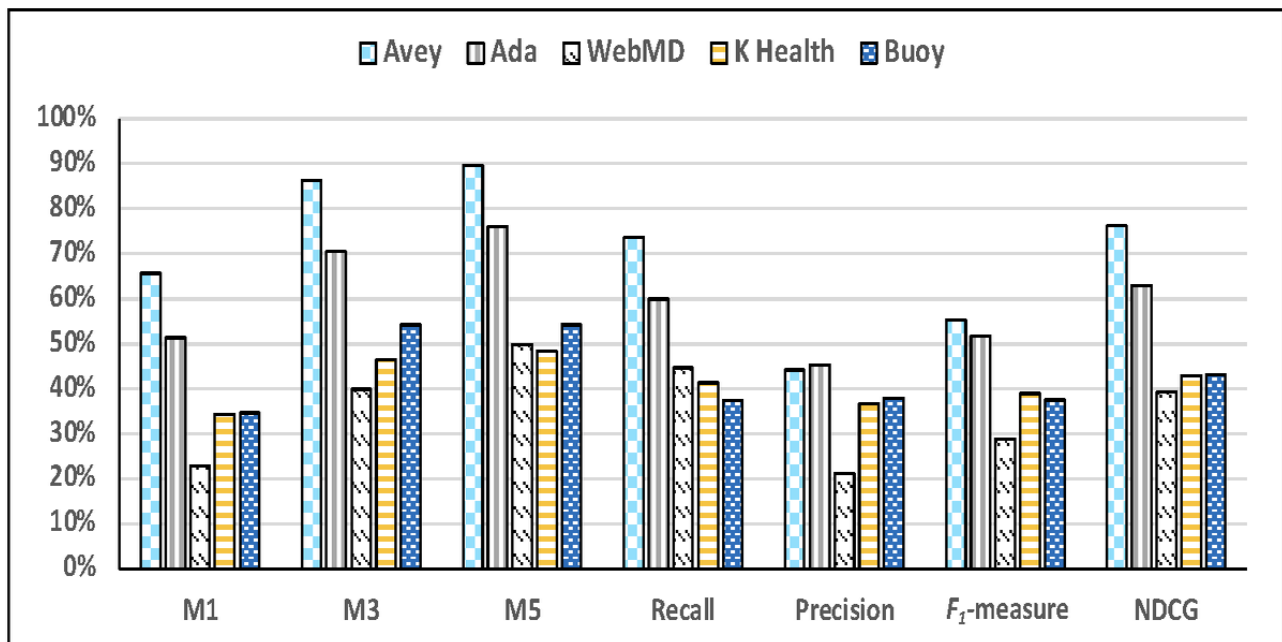


Finally, Figure 5 shows the accuracy results of all the symptom checkers over only the vignettes that resulted in differential diagnoses on every symptom checker (ie, the intersection of successful vignettes with differential diagnoses across all symptom checkers). In this set of results, we excluded Babylon as it failed to produce differential diagnoses for 351 out of the 400 vignettes. As demonstrated in the figure, Avey outperformed Ada, WebMD, K Health, and Buoy, respectively, by averages of 28.1%, 186.9%, 91.5%, and 89.3% using M1; 22.4%, 116.3%, 85.6%, and 59.2% using M3; 18%, 80.1%, 85.7%, and 65.5% using M5; 23%, 64.9%, 78.5%, and 97.1% using recall; 7.2%, 92.7%, 42.2%, and 47.1% using F₁-measure; and 21%, 93.6%,

77.4%, and 76.6% using NDCG. Under average precision, Ada surpassed Avey by an average of 2.4%, whereas Avey outpaced WebMD, K Health, and Buoy by averages of 109.5%, 20.4%, and 16.9%, respectively.

All the combinations of all the results (ie, 45 sets of experiments), including a breakdown between common and less-common diseases, are posted on the internet [27]. In general, we found Avey to be more accurate than the other 5 tested symptom checkers, irrespective of the combination of results; hence, it was chosen to be compared against primary care physicians.

Figure 5. Accuracy results considering only the succeeded vignettes with differential diagnoses across all the symptom checkers. NDCG: Normalized Discounted Cumulative Gain.



Avey Versus Human Doctors

In this section, we present our findings of stage 4. As discussed in the Methods section, we tested the 400 gold-standard vignettes on 3 doctors with an average clinical experience of 16.6 years. Table 4 shows the results of the doctors across all our accuracy metrics. Furthermore, Multimedia Appendix 2 depicts the results of Avey against the average physician, which is the average performance of the 3 physicians. As shown, the human doctors provided average *M1*, *M3*, *M5*, recall, precision, *F₁*-measure, and NDCG of 61.2%, 72.5%, 72.9%, 46.6%, 69.5%, 55.3%, and 61.2%, respectively. In contrast, Avey demonstrated

average *M1*, *M3*, *M5*, recall, precision, *F₁*-measure, and NDCG of 67.5%, 87.3%, 90%, 72.9%, 43.7%, 54.6%, and 76.6%, respectively.

To this end, Avey compared favorably to the considered doctors, yielding inferior performance in terms of precision and *F₁*-measure but a better performance in terms of *M1*, *M3*, *M5*, NDCG, and recall. More precisely, the doctors outperformed Avey by averages of 37.1% and 1.2% using precision and *F₁*-measure, whereas Avey outpaced them by averages of 10.2%, 20.4%, 23.4%, 56.4%, and 25.1% using *M1*, *M3*, *M5*, recall, and NDCG, respectively.

Table 4. Accuracy results (%) of 3 medical doctors (MDs), MD₁, MD₂, and MD₃, with an average experience of 16.6 years.

Doctors	M1	M3	M5	Recall	Precision	<i>F₁</i> -measure	NDCG ^a
MD ₁	49.7	62	62.7	41.2	58.6	48.4	52.2
MD ₂	61.3	67.2	67.5	41.2	78.1	53.9	58
MD ₃	72.5	88.2	88.5	57.3	71.7	63.7	73.5

^aNDCG: Normalized Discounted Cumulative Gain.

Discussion

Principal Findings

In this paper, we capitalized on the standard clinical vignette approach to assess the accuracies of 6 symptom checkers and 3 primary care physicians with an average experience of 16.6 years. We found that Avey is the most accurate among the considered symptom checkers and compares favorably to the 3 involved physicians. For instance, under *M1*, Avey outperforms

the next best-performing symptom checker, namely, Ada, by 24.5% and the worst-performing symptom checker, namely Babylon, by 2968.2%. On average, Avey outperforms the 5 competing symptom checkers by 694.1% using *M1*. In contrast, under *M1*, Avey underperforms the best-performing physician by 6.9% and outperforms the worst-performing one by 35.8%. On average, Avey outperforms the 3 physicians by 13% using *M1*. Table 5 shows the ordering of symptoms and physicians from best-performing to worst-performing.

Table 5. Ordering of symptom checkers and physicians (denoted as MD₁, MD₂, and MD₃) from best-performing to worst-performing symptom checkers and physicians.

Metrics	Descending order (best to worst)	Symptom checkers		Doctors	
		Values, range (%)	Values, SD (%)	Values, range (%)	Values, SD (%)
M1%	MD ₃ , Avey, MD ₂ , Ada, MD ₁ , K Health, Buoy, WebMD, and Babylon	65.3	21	22.8	9
M3%	MD ₃ , Avey, Ada, MD ₂ , MD ₁ , WebMD, Buoy, K Health, and Babylon	84.8	27	26.2	11
M5%	Avey, MD ₃ , Ada, MD ₂ , MD ₁ , WebMD, K Health, Buoy, and Babylon	87.2	27	25.8	11
Average recall	Avey, Ada, MD ₃ , WebMD, MD ₁ and MD ₂ (a tie), K Health, Buoy, and Babylon	70.9	22	16.1	8
Average precision	MD ₃ , MD ₂ , MD ₁ , Ada, Avey, K Health, Buoy, WebMD, and Babylon	40.6	13	19.5	8
Average F ₁ -measure	MD ₃ , Avey, MD ₂ , Ada, MD ₁ , K Health, Buoy and WebMD (a tie), and Babylon	32.9	16	15.3	6
Average ND-CG ^a	Avey, MD ₃ , Ada, MD ₂ , MD ₁ , WebMD, K Health, Buoy, and Babylon	74.2	23	21.3	9

^aNDCCG: Normalized Discounted Cumulative Gain.

Strengths and Limitations

This paper proposed a comprehensive and rigorous experimentation methodology that taps into the standard clinical vignette approach to evaluate symptom checkers and primary care physicians. On the basis of this methodology, we developed and peer reviewed the largest benchmark vignette suite in the domain thus far. A recent study used 200 vignettes and was deemed one of the most comprehensive to date [28]. The work of Semigran et al [29] used 45 vignettes and many studies followed suit [4,7,12,38].

Using this standardized suite, we evaluated the performance of a new AI symptom checker, namely, Avey; 5 popular symptom checkers, namely, Ada, WebMD, K Health, Buoy, and Babylon; and a panel of 3 experienced physicians to put things in perspective and interpret results accordingly. To measure accuracy, we used 7 standard metrics, one of which was leveraged for the first time in literature to quantify the ranking qualities of symptom checkers' and physicians' differential diagnoses. To minimize bias, the 6 symptom checkers were tested by only independent primary care physicians and using only peer-reviewed vignettes.

To facilitate the reproducibility of the study and support future related studies, we made all the peer-reviewed vignettes publicly and freely available on the internet [27]. In addition, we posted on the internet all the reported results (eg, the screenshots of the sessions with symptom checkers and the answers of physicians) on the Benchmark Vignette Suite [27] to establish a standard of full transparency and allow for external cross-validation.

That said, this study lacks an evaluation with real patients and covers only 14 body systems with a limited range of conditions. As pointed out in the Methods section, in the clinical vignette approach, vignettes act as proxies for real patients. The first step in this approach is to standardize these vignettes, which would necessitate an assembly of independent and experienced

physicians to review and approve them. Consequently, if we replace vignettes with real patients, a group of physicians (say, 7, as is the case in this study, for example) is needed to check each patient at the same time and agree by a supermajority vote on their differential diagnosis. This corresponds to standardizing the diagnosis of the patient before she or he is asked to self-diagnose with each symptom checker. Afterward, the diagnoses of the symptom checkers can be matched against the patient's standardized diagnosis and accuracy results can be reported accordingly.

Albeit appealing, the abovementioned approach differs from the standard clinical vignette approach (wherein no vignettes will be involved anymore but actual patients) and is arguably less practical, especially since it suggests checking and diagnosing a vast number of patients, each by a panel of physicians, before testing on symptom checkers. In addition, the cases of the patients should cover enough diseases (eg, as in Table 1), which could drastically increase the pool of patients that needs to be diagnosed by physicians before identifying a representative sample. This may explain why this alternative approach has not been used in any of the accuracy studies of symptom checkers so far, granted that the clinical vignette approach is a standard paradigm, let alone that it is also commonly used for testing the diagnostic abilities of physicians [29].

In any of these approaches, it is important to distinguish between *testers* and *subjects*. For instance, in the abovementioned alternative approach, the patients are the testers of the symptom checkers and the subjects by which the symptom checkers are tested. In contrast, in the clinical vignette approach, the testers are either physicians or laypeople, whereas the subjects are the standardized vignettes. As discussed in the Stage 3: Vignette Testing on Symptom Checkers section, using physicians as testers makes the clinical vignette approach more reliable. This is because symptom checkers may ask questions that hold no answers in the standardized vignettes, making it difficult for

laypeople to answer them appropriately and hard for the community to trust the reported results accordingly.

To this end, 2 research methodologies have been adopted in the literature. One is to dry run a priori by a physician every gold-standard vignette on every considered symptom checker and identify every finding (ie, symptom, etiology, or attribute) that could be asked by these symptom checkers. Subsequently, the physician supplements each vignette with more findings to ensure that laypeople can properly answer any question asked during actual testing. This is the methodology that was used in the seminal work of Semigran et al [11,29].

The second methodology is not to dry run each vignette beforehand on each symptom checker, especially as it might not be possible to fully know what an AI-based symptom checker will ask during actual testing. On the contrary, the methodology suggests standardizing the vignettes in a way that precisely reflects real-life patient cases. Afterward, multiple (to address bias and ensure reliability) independent physicians test the vignettes on each symptom checker. These physicians will then reliably answer any questions about any data not included in the vignettes, thus ensuring the correctness of the approach. This methodology has been shown to be more reliable for conducting accuracy studies [28,38,47]. As such, it was used in most recent state-of-the-art papers [4,28] and, consequently, in ours.

Aside from studying the accuracy of symptom checkers, real patients can be involved in testing the usability of such tools (eg, by using a self-completed questionnaire after self-diagnosing with symptom checkers as in the study by Miller et al [55]). Clearly, this type of study is orthogonal to the accuracy ones and lies outside the scope of this paper.

Finally, we indicate that the physicians that were compared against the symptom checkers in stage 4 (ie, vignette testing on doctors) may not be a representative sample of primary care physicians. Furthermore, our study did not follow a rigorous process to choose symptom checkers and considered only a few of them, which were either new (ie, Avey), popular (ie, Babylon), or performed superiorly in recent studies (ie, Ada, K Health, Buoy, and WebMD).

Comparison With the Wider Literature

Much work, especially recently, has been done to study symptom checkers from different perspectives. It is not possible to do justice to this large body of work in this short paper. As such, we briefly describe some of the most closely related ones, which focus primarily on the accuracy of self-diagnosis.

Semigran et al [29] were the first to study the performance of many symptom checkers across a range of conditions in 2015. They tested 45 vignettes over 23 symptom checkers and discovered that their accuracies vary considerably, with *M1* ranging from 5% to 50% and *M20* (which measures if a symptom checker returns the gold-standard main diagnosis among its top 20 suggested conditions) ranging from 34% to 84%.

Semigran et al [11] published a follow-up paper in 2016 that compared the diagnostic accuracies of physicians against

symptom checkers using the same vignettes in Semigran et al [29]. Results showed that, on average, physicians outperformed symptom checkers (72.1% vs 34.0% along *M1* and 84.3% vs 51.2% along *M3*). However, symptom checkers were more likely to output the gold-standard main diagnosis at the top of their differentials for low-acuity and common vignettes, whereas physicians were more likely to do so for high-acuity and uncommon vignettes.

The 2 studies of Semigran et al [11,29] provided useful insights into the first generation of symptom checkers. However, much has changed from 2015 to 2016. To exemplify, Gilbert et al [28] recently compiled, peer reviewed, and tested 200 vignettes over 8 popular symptom checkers and 7 primary care physicians. As in the study by Semigran et al [29], they found a significant variance in the performance of symptom checkers, but a promise in the accuracy of a new symptom checker named Ada [41]. Ada exhibited accuracies of 49%, 70.5%, and 78% under *M1*, *M3*, and *M5*, respectively.

None of the symptom checkers in the study by Gilbert et al [28] outperformed general practitioners but Ada came close, especially in *M3* and *M5*. The authors of the study by Gilbert et al [28] pointed out that the nature of iterative improvements in software suggests an expected increase in the future performance of symptom checkers, which may at a point in time exceed that of general practitioners. As illustrated in Figure 2, we found that Ada is still largely ahead of the conventional symptom checkers but Avey outperforms it. Furthermore, Avey surpassed a panel of physicians under various accuracy metrics as depicted in Multimedia Appendix 2.

Hill et al [4] evaluated 36 symptom checkers, 8 of which use AI, over 48 vignettes. They showed that accuracy varies considerably across symptom checkers, ranging from 12% to 61% using *M1* and from 30% to 81% using *M10* (where the correct diagnosis appears among the top 10 conditions). They also observed that AI-based symptom checkers outperform rule-based ones (ie, symptom checkers that do not use AI). Akin to Hill et al [4], Ceney et al [12] detected a significant variation in accuracy across 12 symptom checkers, ranging from 22.2% (Caidr [56]) to 72% (Ada) using *M5*.

Many other studies focused on the diagnostic performance of symptom checkers, but only across a limited set of diagnoses [57-68]. For instance, Shen et al [67] evaluated the accuracy of WebMD for ophthalmic diagnoses. Hennemann et al [62] investigated the diagnostic performance of Ada for mental disorders. Nateqi et al [65] validated the accuracies of Symptoma [69], Ada, FindZebra [70], Mediktör [71], Babylon, and Isabel [72] for ear, nose, and throat conditions. Finally, Munsch et al [64] assessed the accuracies of 10 web-based COVID-19 symptom checkers.

From a technical perspective, early AI models for medical diagnosis adopted expert systems [49,73-76]. Subsequent models used probabilistic formulations to account for uncertainty in the diagnostic process [77] and focused on approximate probabilistic inference to optimize for efficiency [78-80].

With the increasing availability of electronic medical records (EMRs), Rotmensch et al [81] used logistic regression, naive

Bayes, and Bayesian networks with noisy OR gates (noisy OR) on EMRs to automatically construct medical knowledge graphs. Miotto et al [82] proposed an EMR-based unsupervised deep learning approach to derive a general-purpose patient representation and facilitate clinical predictive modeling. Ling et al [83] modeled the problem as a sequential decision-making process using deep reinforcement learning. Kannan et al [49] showed that multiclass logistic regression and deep learning models can be effective in generalizing to new patient cases, but with an accuracy caveat concerning the number of diseases that can be incorporated.

Miller et al [55] presented a real-world usability study of Ada over 523 participants (patients) in a South London primary care clinic over a period of 3 months. Approximately all patients (ie, 97.8%) found Ada very easy to use. In addition, 22% of patients aged between 18 and 24 years suggested that using Ada before coming to the clinic would have changed their minds in terms of what care to consider next. Studies of other symptom checkers such as Buoy and Isabel reported high degrees of utility as well [24,84].

Some other work has also explored the triage capabilities of symptom checkers [7,38,84-86]. Studying the utility and triage capabilities of symptom checkers is beyond the scope of this paper and has been set as future work in the Unanswered Questions and Future Research section.

Finally, we note that many survey papers systematically reviewed symptom checkers, made several observations, and identified a few gaps [12,20,23,53,86-91]. For instance, Chambers et al [87] found in 2019 that symptom checkers were much less accurate than physicians. This was observed in this study as well for most of the symptom checkers (see the Results section). Aboueid et al [12] identified knowledge gaps in the literature and recommended producing more research in this area with a focus on accuracy, user experience, regulation, doctor-patient relationship, primary care provider perspectives, and ethics. Finally, some studies [88-90] highlighted various challenges and opportunities in using symptom checkers. They revealed methodological variability in triage and diagnostic accuracies and, thus, urged for more rigorous and standardized evaluations before widespread adoption. In response to this, our work used the standard clinical vignette approach to study the diagnostic accuracies of some commonly used symptom checkers.

Implications for Clinicians and Policy Makers

As pointed out in the Introduction section, a United Kingdom-based study that engaged 1071 patients found that >70% of individuals aged between 18 and 39 years would use

a symptom checker [13]. This study was influential in the United Kingdom health policy circles, whereby it received press attention and prompted responses from National Health Service England and National Health Service X, a United Kingdom government policy unit that develops best practices and national policies for technology in health [55,92]. Given that symptom checkers vary considerably in performance (as demonstrated in the Results section), this paper serves to scientifically inform patients, clinicians, and policy makers about the current accuracies of some of these symptom checkers.

Finally, this study suggests that any external scientific validation of any AI-based medical diagnostic algorithm should be fully transparent and eligible for replication. As a direct translation to this suggestion, we posted all the results of the tested symptom checkers and physicians on the web to allow for cross-verification and study replication. Moreover, we made all peer-reviewed vignettes in our study publicly and freely available. This will not only enable the reproducibility of our study but also further support future related studies, both in academia and industry alike.

Unanswered Questions and Future Research

This paper focused solely on studying the diagnostic accuracies of symptom checkers. Consequently, we set forth 2 complementary future directions, namely, usability and utility. To elaborate, we will first study the usability and acceptability of symptom checkers with real patients. In particular, we will investigate how patients will perceive symptom checkers and interact with them. During this study, we will observe and identify any barrier in the user experience or user interface and language characteristics of such symptom checkers. Finally, we will examine how patients will respond to the output of these symptom checkers and gauge their influence on their subsequent choices for care, especially when it comes to triaging.

Conclusions

In this paper, we proposed an experimentation methodology that taps into the standard clinical vignette approach to evaluate and analyze 6 symptom checkers. To put things in perspective, we further compared the symptom checker that demonstrated the highest performance, namely, Avey against a panel of experienced primary care physicians. Results showed that Avey outperforms the 5 other considered symptom checkers, namely, Ada, K Health, Buoy, Babylon, and WebMD by a large margin and compares favorably to the participating physicians. The nature of iterative improvements in software and the fast pace of advancements in AI suggest an accelerated increase in the future performance of such symptom checkers.

Acknowledgments

The vignette setting was carried out with the help of the following independent and experienced physicians: Dr Azmi Qudsi, Dr Doaa Eisa, and Dr Muna Yousif. Vignette review (ie, vignette standardization, or stage 2 of our experimentation methodology) was carried out by the following independent and experienced physicians: Dr Zaid Abu Saleh, Dr Odai Al-Batsh, Dr Ahmad Alowaidat, Dr Tamara Altawara, Dr Arwa Khashan, Dr Muna Darmach, and Dr Nour Essale. Vignette testing on symptom checkers (ie, stage 3 of our experimentation methodology) was carried out by the following independent and experienced physicians: Dr Maram Alsmairat, Dr Muna Darmach, and Dr Ahmad Kakakan. Vignette testing on doctors (ie, stage 4 of our experimentation

methodology) was carried out by the following independent and experienced physicians: Dr Mohammad Almadani, Dr Tala Hamouri, and Dr Noor Jodeh.

Data Availability

All our gold-standard vignettes are made publicly and freely available [93] to enable the reproducibility of this work. In addition, all the outputs of the symptom checkers and physicians are posted at the same site to allow for external cross-validation. Finally, the results of all our 45 sets of experiments are published [94] to establish a standard of full transparency.

Disclaimer

The guarantor (MH) affirms that the manuscript is an honest, accurate, and transparent account of the study being reported; that no important aspects of the study have been omitted; and that any discrepancies from the study as planned (and, if relevant, registered) have been explained.

Authors' Contributions

The first author (MH) conceived the study, designed the experimentation methodology, and supervised the project. The second author (SD) coordinated the work within and across the project stages (eg, coordination of vignette creation, vignette standardization, vignette testing on symptom checkers, and vignette testing on doctors). The first author (MH) conducted the literature review and documentation. The second, third, and fourth authors (SD, MD, and SA) created the vignettes and verified the testing results. The third and the fifth authors (MD and SS) carried out results compilation and summarization. The third and the fifth authors (MD and SS) carried out data analysis and interpretation. The sixth author (YK) developed the web portal for streamlining the processes of reviewing, standardizing, and testing the vignettes. The fifth author (SS) maintained Avey's software and provided technical support. The first author (MH) wrote the paper. All authors (MH, SD, MD, SA, SS, and YK) reviewed and commented on drafts of the paper. The first author (MH) provided administrative support and is the guarantor for this work.

Conflicts of Interest

All authors have completed The International Committee of Medical Journal Editors uniform disclosure form [95]. All authors are employees of Avey Inc, which is the manufacturer of Avey (see authors' affiliations). The first author is the founder and CEO of Avey Inc and holds equity in it. The authors have no support from any organization for the submitted work; no financial relationships with any organizations that might have interests in the submitted work; and no other relationships or activities that could appear to have influenced the submitted work.

Multimedia Appendix 1

The alignment of our methodology with the recommended requirements of pursuing the clinical vignette approach.

[DOCX File, 11 KB - [ai_v3i1e46875_app1.docx](#)]

Multimedia Appendix 2

Accuracy results of Avey versus 3 medical doctors (MDs), on average (ie, average MD). NDCG: Normalized Discounted Cumulative Gain.

[PNG File, 88 KB - [ai_v3i1e46875_app2.png](#)]

References

1. Morahan-Martin JM. How internet users find, evaluate, and use online health information: a cross-cultural review. *Cyberpsychol Behav* 2004 Oct;7(5):497-510. [doi: [10.1089/cpb.2004.7.497](#)] [Medline: [15667044](#)]
2. Wyatt JC. Fifty million people use computerised self triage. *BMJ* 2015 Jul 08;351:h3727. [doi: [10.1136/bmj.h3727](#)] [Medline: [26156750](#)]
3. Cheng C, Dunn M. Health literacy and the internet: a study on the readability of Australian online health information. *Aust N Z J Public Health* 2015 Aug;39(4):309-314 [FREE Full text] [doi: [10.1111/1753-6405.12341](#)] [Medline: [25716142](#)]
4. Hill MG, Sim M, Mills B. The quality of diagnosis and triage advice provided by free online symptom checkers and apps in Australia. *Med J Aust* 2020 Jun 11;212(11):514-519. [doi: [10.5694/mja2.50600](#)] [Medline: [32391611](#)]
5. Levine DM, Mehrotra A. Assessment of diagnosis and triage in validated case vignettes among nonphysicians before and after internet search. *JAMA Netw Open* 2021 Mar 01;4(3):e213287 [FREE Full text] [doi: [10.1001/jamanetworkopen.2021.3287](#)] [Medline: [33779741](#)]
6. Martin SS, Quayle E, Schultz S, Fashanu OE, Wang J, Saheed MO, et al. A randomized controlled trial of online symptom searching to inform patient generated differential diagnoses. *NPJ Digit Med* 2019;2:110 [FREE Full text] [doi: [10.1038/s41746-019-0183-0](#)] [Medline: [31728417](#)]

7. Schmieding ML, Mörgeli R, Schmieding MA, Feufel MA, Balzer F. Benchmarking triage capability of symptom checkers against that of medical laypersons: survey study. *J Med Internet Res* 2021 Mar 10;23(3):e24475 [FREE Full text] [doi: [10.2196/24475](https://doi.org/10.2196/24475)] [Medline: [33688845](https://pubmed.ncbi.nlm.nih.gov/33688845/)]
8. Norman B. Don't google it. Vimeo. URL: <https://vimeo.com/115097884> [accessed 2022-01-08]
9. Larimer S. Can this ad campaign get people in Belgium to stop Googling their symptoms? *Washington Post*. 2014 Nov 11. URL: <https://www.washingtonpost.com/news/to-your-health/wp/2014/11/11/can-this-ad-campaign-get-people-in-belgium-to-stop-googling-their-symptoms/> [accessed 2022-01-08]
10. Aboueid S, Meyer S, Wallace JR, Mahajan S, Chaurasia A. Young adults' perspectives on the use of symptom checkers for self-triage and self-diagnosis: qualitative study. *JMIR Public Health Surveill* 2021 Jan 06;7(1):e22637 [FREE Full text] [doi: [10.2196/22637](https://doi.org/10.2196/22637)] [Medline: [33404515](https://pubmed.ncbi.nlm.nih.gov/33404515/)]
11. Semigran HL, Levine DM, Nundy S, Mehrotra A. Comparison of physician and computer diagnostic accuracy. *JAMA Intern Med* 2016 Dec 01;176(12):1860-1861. [doi: [10.1001/jamainternmed.2016.6001](https://doi.org/10.1001/jamainternmed.2016.6001)] [Medline: [27723877](https://pubmed.ncbi.nlm.nih.gov/27723877/)]
12. Ceney A, Tolond S, Glowinski A, Marks B, Swift S, Palser T. Accuracy of online symptom checkers and the potential impact on service utilisation. *PLoS One* 2021 Jul 15;16(7):e0254088 [FREE Full text] [doi: [10.1371/journal.pone.0254088](https://doi.org/10.1371/journal.pone.0254088)] [Medline: [34265845](https://pubmed.ncbi.nlm.nih.gov/34265845/)]
13. Using technology to ease the burden on primary care. Healthwatch Enfield. URL: https://www.healthwatchenfield.co.uk/sites/healthwatchenfield.co.uk/files/Report_UsingTechnologyToEaseTheBurdenOnPrimaryCare.pdf [accessed 2022-01-08]
14. Meyer AN, Giardina TD, Spitzmueller C, Shahid U, Scott TM, Singh H. Patient perspectives on the usefulness of an artificial intelligence-assisted symptom checker: cross-sectional survey study. *J Med Internet Res* 2020 Jan 30;22(1):e14679 [FREE Full text] [doi: [10.2196/14679](https://doi.org/10.2196/14679)] [Medline: [32012052](https://pubmed.ncbi.nlm.nih.gov/32012052/)]
15. Access NHS clinicians 24/7. Babylon GP at Hand. URL: <https://www.gpathand.nhs.uk/our-nhs-service> [accessed 2022-01-08]
16. healthdirect symptom checker. Healthdirect Australia. URL: <https://about.healthdirect.gov.au/healthdirect-symptom-checker> [accessed 2022-01-08]
17. Spoelman WA, Bonten TN, de Waal MW, Drenthen T, Smelee IJ, Nielen MM, et al. Effect of an evidence-based website on healthcare usage: an interrupted time-series study. *BMJ Open* 2016 Nov 09;6(11):e013166 [FREE Full text] [doi: [10.1136/bmjopen-2016-013166](https://doi.org/10.1136/bmjopen-2016-013166)] [Medline: [28186945](https://pubmed.ncbi.nlm.nih.gov/28186945/)]
18. Aboueid S, Liu RH, Desta BN, Chaurasia A, Ebrahim S. The use of artificially intelligent self-diagnosing digital platforms by the general public: scoping review. *JMIR Med Inform* 2019 May 01;7(2):e13445 [FREE Full text] [doi: [10.2196/13445](https://doi.org/10.2196/13445)] [Medline: [31042151](https://pubmed.ncbi.nlm.nih.gov/31042151/)]
19. Brownlee S, Chalkidou K, Doust J, Elshaug AG, Glasziou P, Heath I, et al. Evidence for overuse of medical services around the world. *The Lancet* 2017 Jul;390(10090):156-168. [doi: [10.1016/s0140-6736\(16\)32585-5](https://doi.org/10.1016/s0140-6736(16)32585-5)]
20. Morgan DJ, Dhruva SS, Wright SM, Korenstein D. 2016 update on medical overuse: a systematic review. *JAMA Intern Med* 2016 Nov 01;176(11):1687-1692 [FREE Full text] [doi: [10.1001/jamainternmed.2016.5381](https://doi.org/10.1001/jamainternmed.2016.5381)] [Medline: [27654002](https://pubmed.ncbi.nlm.nih.gov/27654002/)]
21. Unnecessary care in Canada. Canadian Institute for Health Information. 2017 Apr. URL: <https://www.cihi.ca/sites/default/files/document/choosing-wisely-baseline-report-en-web.pdf> [accessed 2022-01-08]
22. Aboueid S, Meyer SB, Wallace JR, Mahajan S, Nur T, Chaurasia A. Use of symptom checkers for COVID-19-related symptoms among university students: a qualitative study. *BMJ Innov* 2021 Apr;7(2):253-260. [doi: [10.1136/bmjinnov-2020-000498](https://doi.org/10.1136/bmjinnov-2020-000498)] [Medline: [34192014](https://pubmed.ncbi.nlm.nih.gov/34192014/)]
23. Akbar S, Coiera E, Magrabi F. Safety concerns with consumer-facing mobile health applications and their consequences: a scoping review. *J Am Med Inform Assoc* 2020 Feb 01;27(2):330-340 [FREE Full text] [doi: [10.1093/jamia/ocz175](https://doi.org/10.1093/jamia/ocz175)] [Medline: [31599936](https://pubmed.ncbi.nlm.nih.gov/31599936/)]
24. Fraser H, Coiera E, Wong D. Safety of patient-facing digital symptom checkers. *The Lancet* 2018 Nov;392(10161):2263-2264. [doi: [10.1016/s0140-6736\(18\)32819-8](https://doi.org/10.1016/s0140-6736(18)32819-8)]
25. Kasteleyn MJ, Versluis A, van Peet P, Kirk UB, van Daltsen J, Meijer E, et al. SERIES: eHealth in primary care. Part 5: a critical appraisal of five widely used eHealth applications for primary care - opportunities and challenges. *Eur J Gen Pract* 2021 Dec;27(1):248-256 [FREE Full text] [doi: [10.1080/13814788.2021.1962845](https://doi.org/10.1080/13814788.2021.1962845)] [Medline: [34432601](https://pubmed.ncbi.nlm.nih.gov/34432601/)]
26. Painter A, Hayhoe B, Riboli-Sasco E, El-Osta A. Online symptom checkers: recommendations for a vignette-based clinical evaluation standard. *J Med Internet Res* 2022 Oct 26;24(10):e37408 [FREE Full text] [doi: [10.2196/37408](https://doi.org/10.2196/37408)] [Medline: [36287594](https://pubmed.ncbi.nlm.nih.gov/36287594/)]
27. Avey's Benchmark Vignette Suite. Avey. URL: <https://avey.ai/research/avey-accurate-ai-algorithm/benchmark-vignette-suite> [accessed 2024-04-02]
28. Gilbert S, Mehl A, Baluch A, Cawley C, Challiner J, Fraser H, et al. How accurate are digital symptom assessment apps for suggesting conditions and urgency advice? A clinical vignettes comparison to GPs. *BMJ Open* 2020 Dec 16;10(12):e040269 [FREE Full text] [doi: [10.1136/bmjopen-2020-040269](https://doi.org/10.1136/bmjopen-2020-040269)] [Medline: [33328258](https://pubmed.ncbi.nlm.nih.gov/33328258/)]
29. Semigran HL, Linder JA, Gidengil C, Mehrotra A. Evaluation of symptom checkers for self diagnosis and triage: audit study. *BMJ* 2015 Jul 08;351:h3480 [FREE Full text] [doi: [10.1136/bmj.h3480](https://doi.org/10.1136/bmj.h3480)] [Medline: [26157077](https://pubmed.ncbi.nlm.nih.gov/26157077/)]
30. Step 2 CK. United States Medical Licensing Examination. URL: <https://www.usmle.org/step-exams/step-2-ck> [accessed 2022-02-05]

31. Firth JD, Newman M. MRCP Part 1 Self-Assessment: Medical Masterclass Questions and Explanatory Answers. Boca Raton, FL: CRC Press; 2008.
32. Knutson D. Family Medicine: PreTest Self-assessment and Review. New York, NY: McGraw-Hill Medical; 2012.
33. In-training examination. American Board of Family Medicine. URL: <https://www.theabfm.org/become-certified/acgme-program/in-training-examination> [accessed 2024-04-02]
34. American Academy of Pediatrics. URL: <https://www.aap.org/> [accessed 2022-02-05]
35. 100 Cases book series. Routledge. URL: <https://www.routledge.com/100-Cases/book-series/CRCONEHUNCAS> [accessed 2022-02-05]
36. Tallia AF, Scherger JE, Dickey N. Swanson's Family Medicine Review. Amsterdam, The Netherlands: Elsevier; 2021.
37. Wilkinson IB, Raine T, Wiles K, Goodhart A, Hall C, O'Neill H. Oxford Handbook of Clinical Medicine. Oxford, UK: Oxford University Press; Jul 2017.
38. Baker A, Perov Y, Middleton K, Baxter J, Mullarkey D, Sangar D, et al. A comparison of artificial intelligence and human doctors for the purpose of triage and diagnosis. *Front Artif Intell* 2020 Nov 30;3:543405 [FREE Full text] [doi: [10.3389/frai.2020.543405](https://doi.org/10.3389/frai.2020.543405)] [Medline: [33733203](https://pubmed.ncbi.nlm.nih.gov/33733203/)]
39. Avey. URL: <https://avey.ai/research> [accessed 2024-04-02]
40. Avey app. Avey. URL: <https://avey.ai/> [accessed 2024-04-02]
41. Health. Powered by Ada. Ada. URL: <https://ada.com/> [accessed 2022-01-07]
42. K Health: 24/7 access to high-quality medicine. K Health. URL: <https://khealth.com/> [accessed 2022-01-07]
43. Buoy health: check symptom and find the right care. Buoy Health. URL: <https://www.buoyhealth.com/> [accessed 2022-01-07]
44. Babylon Healthcare. URL: <https://www.babylonhealth.com/> [accessed 2022-01-07]
45. WebMD. URL: <https://www.webmd.com/> [accessed 2022-01-07]
46. Coiera E, Ammenwerth E, Georgiou A, Magrabi F. Does health informatics have a replication crisis? *J Am Med Inform Assoc* 2018 Aug 01;25(8):963-968 [FREE Full text] [doi: [10.1093/jamia/ocy028](https://doi.org/10.1093/jamia/ocy028)] [Medline: [29669066](https://pubmed.ncbi.nlm.nih.gov/29669066/)]
47. Jungmann SM, Klan T, Kuhn S, Jungmann F. Accuracy of a chatbot (Ada) in the diagnosis of mental disorders: comparative case study with lay and expert users. *JMIR Form Res* 2019 Oct 29;3(4):e13863 [FREE Full text] [doi: [10.2196/13863](https://doi.org/10.2196/13863)] [Medline: [31663858](https://pubmed.ncbi.nlm.nih.gov/31663858/)]
48. Moreno Barriga E, Pueyo Ferrer I, Sánchez Sánchez M, Martín Baranera M, Masip Utset J. [A new artificial intelligence tool for assessing symptoms in patients seeking emergency department care: the Mediktor application]. *Emergencias* 2017;29(6):391-396 [FREE Full text] [Medline: [29188913](https://pubmed.ncbi.nlm.nih.gov/29188913/)]
49. Kannan A, Fries JA, Kramer E, Chen JJ, Shah N, Amatriain X. The accuracy vs. coverage trade-off in patient-facing diagnosis models. *AMIA Jt Summits Transl Sci Proc* 2020 May 30;2020:298-307 [FREE Full text] [Medline: [32477649](https://pubmed.ncbi.nlm.nih.gov/32477649/)]
50. Järvelin K, Kekäläinen J. Cumulated gain-based evaluation of IR techniques. *ACM Trans Inf Syst* 2002 Oct 01;20(4):422-446. [doi: [10.1145/582415.582418](https://doi.org/10.1145/582415.582418)]
51. Wang Y, Wang L, Li Y, He D, Chen W, Liu TY. A theoretical analysis of NDCG ranking measures. In: *Proceedings of Machine Learning Research* 2013. 2013 Presented at: PMLR 2013; April 29-May 1, 2013; Scottsdale, AZ.
52. Office for Human Research Protections. US Department of Health and Human Services. URL: <https://www.hhs.gov/ohrp/index.html> [accessed 2024-04-02]
53. Hirosawa T, Harada Y, Yokose M, Sakamoto T, Kawamura R, Shimizu T. Diagnostic accuracy of differential-diagnosis lists generated by generative pretrained transformer 3 chatbot for clinical vignettes with common chief complaints: a pilot study. *Int J Environ Res Public Health* 2023 Feb 15;20(4):3378 [FREE Full text] [doi: [10.3390/ijerph20043378](https://doi.org/10.3390/ijerph20043378)] [Medline: [36834073](https://pubmed.ncbi.nlm.nih.gov/36834073/)]
54. El-Osta A, Webber I, Alaa A, Bagkeris E, Mian S, Taghavi Azar Sharabiani M, et al. What is the suitability of clinical vignettes in benchmarking the performance of online symptom checkers? An audit study. *BMJ Open* 2022 Apr 27;12(4):e053566 [FREE Full text] [doi: [10.1136/bmjopen-2021-053566](https://doi.org/10.1136/bmjopen-2021-053566)] [Medline: [35477872](https://pubmed.ncbi.nlm.nih.gov/35477872/)]
55. Miller S, Gilbert S, Virani V, Wicks P. Patients' utilization and perception of an artificial intelligence-based symptom assessment and advice technology in a British primary care waiting room: exploratory pilot study. *JMIR Hum Factors* 2020 Jul 10;7(3):e19713 [FREE Full text] [doi: [10.2196/19713](https://doi.org/10.2196/19713)] [Medline: [32540836](https://pubmed.ncbi.nlm.nih.gov/32540836/)]
56. Squarespace. URL: <https://caidr.squarespace.com/> [accessed 2022-01-08]
57. Berry AC, Berry NA, Wang B, Mulekar MS, Melvin A, Battiola RJ, et al. Use of online symptom checkers to delineate the ever-elusive GERD versus non-GERD cough. *Clin Respir J* 2018 Dec;12(12):2683-2685 [FREE Full text] [doi: [10.1111/crj.12966](https://doi.org/10.1111/crj.12966)] [Medline: [30260573](https://pubmed.ncbi.nlm.nih.gov/30260573/)]
58. Berry AC, Berry NA, Wang B, Mulekar MS, Melvin A, Battiola RJ, et al. Symptom checkers versus doctors: a prospective, head - to - head comparison for cough. *Clin Respir J* 2020 Apr;14(4):413-415. [doi: [10.1111/crj.13135](https://doi.org/10.1111/crj.13135)] [Medline: [31860762](https://pubmed.ncbi.nlm.nih.gov/31860762/)]
59. Bisson LJ, Komm JT, Bernas GA, Fineberg MS, Marzo JM, Rauh MA, et al. Accuracy of a computer-based diagnostic program for ambulatory patients with knee pain. *Am J Sports Med* 2014 Oct;42(10):2371-2376. [doi: [10.1177/0363546514541654](https://doi.org/10.1177/0363546514541654)] [Medline: [25073597](https://pubmed.ncbi.nlm.nih.gov/25073597/)]
60. Ćirković A. Evaluation of four artificial intelligence-assisted self-diagnosis apps on three diagnoses: two-year follow-up study. *J Med Internet Res* 2020 Dec 04;22(12):e18097 [FREE Full text] [doi: [10.2196/18097](https://doi.org/10.2196/18097)] [Medline: [33275113](https://pubmed.ncbi.nlm.nih.gov/33275113/)]

61. Farmer SE, Bernardotto M, Singh V. How good is internet self-diagnosis of ENT symptoms using boots WebMD symptom checker? *Clin Otolaryngol* 2011 Oct;36(5):517-518. [doi: [10.1111/j.1749-4486.2011.02375.x](https://doi.org/10.1111/j.1749-4486.2011.02375.x)] [Medline: [22032458](https://pubmed.ncbi.nlm.nih.gov/22032458/)]
62. Hennemann S, Kuhn S, Withhöft M, Jungmann SM. Diagnostic performance of an app-based symptom checker in mental disorders: comparative study in psychotherapy outpatients. *JMIR Ment Health* 2022 Jan 31;9(1):e32832 [FREE Full text] [doi: [10.2196/32832](https://doi.org/10.2196/32832)] [Medline: [35099395](https://pubmed.ncbi.nlm.nih.gov/35099395/)]
63. Luger TM, Houston TK, Suls J. Older adult experience of online diagnosis: results from a scenario-based think-aloud protocol. *J Med Internet Res* 2014 Jan 16;16(1):e16 [FREE Full text] [doi: [10.2196/jmir.2924](https://doi.org/10.2196/jmir.2924)] [Medline: [24434479](https://pubmed.ncbi.nlm.nih.gov/24434479/)]
64. Munsch N, Martin A, Gruarin S, Nateqi J, Abdarrahmane I, Weingartner-Ortner R, et al. Diagnostic accuracy of web-based COVID-19 symptom checkers: comparison study. *J Med Internet Res* 2020 Oct 06;22(10):e21299 [FREE Full text] [doi: [10.2196/21299](https://doi.org/10.2196/21299)] [Medline: [33001828](https://pubmed.ncbi.nlm.nih.gov/33001828/)]
65. Nateqi J, Lin S, Krobath H, Gruarin S, Lutz T, Dvorak T, et al. [From symptom to diagnosis-symptom checkers re-evaluated: are symptom checkers finally sufficient and accurate to use? An update from the ENT perspective]. *HNO* 2019 May;67(5):334-342. [doi: [10.1007/s00106-019-0666-y](https://doi.org/10.1007/s00106-019-0666-y)] [Medline: [30993374](https://pubmed.ncbi.nlm.nih.gov/30993374/)]
66. Poote AE, French DP, Dale J, Powell J. A study of automated self-assessment in a primary care student health centre setting. *J Telemed Telecare* 2014 Apr;20(3):123-127. [doi: [10.1177/1357633X14529246](https://doi.org/10.1177/1357633X14529246)] [Medline: [24643948](https://pubmed.ncbi.nlm.nih.gov/24643948/)]
67. Shen C, Nguyen M, Gregor A, Isaza G, Beattie A. Accuracy of a popular online symptom checker for ophthalmic diagnoses. *JAMA Ophthalmol* 2019 Jun 01;137(6):690-692. [doi: [10.1001/jamaophthalmol.2019.0571](https://doi.org/10.1001/jamaophthalmol.2019.0571)] [Medline: [30973602](https://pubmed.ncbi.nlm.nih.gov/30973602/)]
68. Yoshida Y, Thomas Clark G. Accuracy of online symptom checkers for diagnosis of orofacial pain and oral medicine disease. *J Prosthodont Res* 2021 Jun 30;65(2):186-190 [FREE Full text] [doi: [10.2186/jpr.JPOR_2019_499](https://doi.org/10.2186/jpr.JPOR_2019_499)] [Medline: [32938875](https://pubmed.ncbi.nlm.nih.gov/32938875/)]
69. Digital health assistant and symptom checker. Symptoma. URL: <https://www.symptoma.com/> [accessed 2022-03-19]
70. FindZebra. URL: <https://www.findzebra.com/> [accessed 2022-03-19]
71. Mediktor. URL: <https://www.mediktor.com/en-us> [accessed 2022-03-19]
72. Isabel - the symptom checker doctors use and trust. Isabel. URL: <https://symptomchecker.isabelhealthcare.com/> [accessed 2022-01-08]
73. Barnett GO, Cimino JJ, Hupp JA, Hoffer EP. DXplain. An evolving diagnostic decision-support system. *JAMA* 1987 Jul 03;258(1):67-74. [doi: [10.1001/jama.258.1.67](https://doi.org/10.1001/jama.258.1.67)] [Medline: [3295316](https://pubmed.ncbi.nlm.nih.gov/3295316/)]
74. Shortliffe EH, Buchanan BG. Rule-based Expert Systems: The MYCIN Experiments of the Stanford Heuristic Programming Project. San Francisco, CA: Addison-Wesley Publishing Company; 1984.
75. Jaakkola TS, Jordan ML. Variational probabilistic inference and the QMR-DT network. *J Artif Intell Res* 1999 May 01;10(1999):291-322. [doi: [10.1613/jair.583](https://doi.org/10.1613/jair.583)]
76. Lauritzen SL, Spiegelhalter DJ. Local computations with probabilities on graphical structures and their application to expert systems. *J R Stat Soc Series B Stat Methodol* 1988;50(2):157-194. [doi: [10.1111/j.2517-6161.1988.tb01721.x](https://doi.org/10.1111/j.2517-6161.1988.tb01721.x)]
77. Miller RA, Pople HEJ, Myers JD. INTERNIST-I, an experimental computer-based diagnostic consultant for general internal medicine. In: Reggia JA, Tuhim S, editors. Computer-Assisted Medical Decision Making. New York, NY: Springer; 1985.
78. Quaid M. Recognition networks for approximate inference in BN20 networks. arXiv Preprint posted online January 10, 2013 [FREE Full text]
79. Rassinoux AM, Miller RA, Baud RH, Scherrer JR. Modeling principles for QMR medical findings. *Proc AMIA Annu Fall Symp* 1996:264-268 [FREE Full text] [Medline: [8947669](https://pubmed.ncbi.nlm.nih.gov/8947669/)]
80. Shwe M, Cooper G. An empirical analysis of likelihood-weighting simulation on a large, multiply connected medical belief network. *Comput Biomed Res* 1991 Oct;24(5):453-475. [doi: [10.1016/0010-4809\(91\)90020-w](https://doi.org/10.1016/0010-4809(91)90020-w)] [Medline: [1743005](https://pubmed.ncbi.nlm.nih.gov/1743005/)]
81. Rotmensch M, Halpern Y, Tlimat A, Horng S, Sontag D. Learning a health knowledge graph from electronic medical records. *Sci Rep* 2017 Jul 20;7(1):5994 [FREE Full text] [doi: [10.1038/s41598-017-05778-z](https://doi.org/10.1038/s41598-017-05778-z)] [Medline: [28729710](https://pubmed.ncbi.nlm.nih.gov/28729710/)]
82. Miotto R, Li L, Kidd BA, Dudley JT. Deep patient: an unsupervised representation to predict the future of patients from the electronic health records. *Sci Rep* 2016 May 17;6:26094 [FREE Full text] [doi: [10.1038/srep26094](https://doi.org/10.1038/srep26094)] [Medline: [27185194](https://pubmed.ncbi.nlm.nih.gov/27185194/)]
83. Ling Y, Hasan SA, Datla V, Qadir A, Lee K, Liu J, et al. Diagnostic inferencing via improving clinical concept extraction with deep reinforcement learning: a preliminary study. In: Proceedings of the 2nd Machine Learning for Healthcare Conference. 2017 Presented at: PMLR 68; August 18-19, 2017; Boston, MA.
84. Winn AN, Somai M, Fergestrom N, Crotty BH. Association of use of online symptom checkers with patients' plans for seeking care. *JAMA Netw Open* 2019 Dec 02;2(12):e1918561 [FREE Full text] [doi: [10.1001/jamanetworkopen.2019.18561](https://doi.org/10.1001/jamanetworkopen.2019.18561)] [Medline: [31880791](https://pubmed.ncbi.nlm.nih.gov/31880791/)]
85. Mansab F, Bhatti S, Goyal D. Reliability of COVID-19 symptom checkers as national triage tools: an international case comparison study. *BMJ Health Care Inform* 2021 Oct;28(1):e100448 [FREE Full text] [doi: [10.1136/bmjhci-2021-100448](https://doi.org/10.1136/bmjhci-2021-100448)] [Medline: [34663637](https://pubmed.ncbi.nlm.nih.gov/34663637/)]
86. Wallace W, Chan C, Chidambaram S, Hanna L, Iqbal FM, Acharya A, et al. The diagnostic and triage accuracy of digital and online symptom checker tools: a systematic review. *NPJ Digit Med* 2022 Aug 17;5(1):118 [FREE Full text] [doi: [10.1038/s41746-022-00667-w](https://doi.org/10.1038/s41746-022-00667-w)] [Medline: [35977992](https://pubmed.ncbi.nlm.nih.gov/35977992/)]

87. Chambers D, Cantrell AJ, Johnson M, Preston L, Baxter SK, Booth A, et al. Digital and online symptom checkers and health assessment/triage services for urgent health problems: systematic review. *BMJ Open* 2019 Aug 01;9(8):e027743 [FREE Full text] [doi: [10.1136/bmjopen-2018-027743](https://doi.org/10.1136/bmjopen-2018-027743)] [Medline: [31375610](https://pubmed.ncbi.nlm.nih.gov/31375610/)]
88. Pairon A, Philips H, Verhoeven V. A scoping review on the use and usefulness of online symptom checkers and triage systems: how to proceed? *Front Med (Lausanne)* 2023 Jan 06;9:1040926 [FREE Full text] [doi: [10.3389/fmed.2022.1040926](https://doi.org/10.3389/fmed.2022.1040926)] [Medline: [36687416](https://pubmed.ncbi.nlm.nih.gov/36687416/)]
89. Riboli-Sasco E, El-Osta A, Alaa A, Webber I, Karki M, El Asmar ML, et al. Triage and diagnostic accuracy of online symptom checkers: systematic review. *J Med Internet Res* 2023 Jun 02;25:e43803 [FREE Full text] [doi: [10.2196/43803](https://doi.org/10.2196/43803)] [Medline: [37266983](https://pubmed.ncbi.nlm.nih.gov/37266983/)]
90. Gottlieb K, Petersson G. Limited evidence of benefits of patient operated intelligent primary care triage tools: findings of a literature review. *BMJ Health Care Inform* 2020 May;27(1):e100114 [FREE Full text] [doi: [10.1136/bmjhci-2019-100114](https://doi.org/10.1136/bmjhci-2019-100114)] [Medline: [32385041](https://pubmed.ncbi.nlm.nih.gov/32385041/)]
91. Harada T, Miyagami T, Kunitomo K, Shimizu T. Clinical decision support systems for diagnosis in primary care: a scoping review. *Int J Environ Res Public Health* 2021 Aug 10;18(16):8435 [FREE Full text] [doi: [10.3390/ijerph18168435](https://doi.org/10.3390/ijerph18168435)] [Medline: [34444182](https://pubmed.ncbi.nlm.nih.gov/34444182/)]
92. Torjesen I. Patients find GP online services "cumbersome," survey finds. *BMJ* 2019 Jul 22;366:l4800. [doi: [10.1136/bmj.l4800](https://doi.org/10.1136/bmj.l4800)] [Medline: [31331913](https://pubmed.ncbi.nlm.nih.gov/31331913/)]
93. Hammoud M, Douglas S, Darmach M, Sanyal S, Alawneh A, Kanbour Y. Evaluating the accuracy of a novel artificial intelligence based symptom checker: a clinical vignettes study: vignette suite and screenshots. Figshare. URL: <https://tinyurl.com/bdh4syvf> [accessed 2024-04-03]
94. Evaluating the accuracy of a novel artificial intelligence based symptom checker: a clinical vignettes study : results document. Figshare. URL: <https://tinyurl.com/45j8atf8> [accessed 2024-04-03]
95. Disclosure of interest (updated February 2021). International Committee of Medical Journal Editors. URL: <https://www.icmje.org/disclosure-of-interest/> [accessed 2024-04-03]

Abbreviations

AI: artificial intelligence

DCG: Discounted Cumulative Gain

EMR: electronic medical record

NDCG: Normalized Discounted Cumulative Gain

Edited by K El Emam, B Malin; submitted 28.02.23; peer-reviewed by B Meskó, S Aboueid; comments to author 31.03.23; revised version received 15.06.23; accepted 02.03.24; published 29.04.24.

Please cite as:

Hammoud M, Douglas S, Darmach M, Alawneh S, Sanyal S, Kanbour Y

Evaluating the Diagnostic Performance of Symptom Checkers: Clinical Vignette Study

JMIR AI 2024;3:e46875

URL: <https://ai.jmir.org/2024/1/e46875>

doi: [10.2196/46875](https://doi.org/10.2196/46875)

PMID: [38875676](https://pubmed.ncbi.nlm.nih.gov/38875676/)

©Mohammad Hammoud, Shahd Douglas, Mohamad Darmach, Sara Alawneh, Swapnendu Sanyal, Youssef Kanbour. Originally published in *JMIR AI* (<https://ai.jmir.org>), 29.04.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in *JMIR AI*, is properly cited. The complete bibliographic information, a link to the original publication on <https://www.ai.jmir.org/>, as well as this copyright and license information must be included.

Original Paper

Near Real-Time Syndromic Surveillance of Emergency Department Triage Texts Using Natural Language Processing: Case Study in Febrile Convulsion Detection

Sedigh Khademi^{1,2}, PhD; Christopher Palmer², Dip Bus Comp; Muhammad Javed², PhD; Gerardo Luis Dimaguila^{1,2,3}, PhD; Hazel Clothier^{1,2,3,4}, PhD; Jim Buttery^{1,2,3,5}, PhD, Prof Dr Med; Jim Black^{4,6}, MD, PhD

¹Department of Paediatrics, University of Melbourne, Melbourne, Australia

²Health Informatics Group, Centre for Health Analytics, Melbourne Children's Campus, Melbourne, Australia

³SAEFVIC, Murdoch Children's Research Institute, Melbourne, Australia

⁴Melbourne School of Population and Global Health, University of Melbourne, Melbourne, Australia

⁵Infectious Diseases, Royal Children's Hospital, Melbourne, Australia

⁶Department of Health, State Government of Victoria, Melbourne, Australia

Corresponding Author:

Sedigh Khademi, PhD

Department of Paediatrics

University of Melbourne

Grattan Street

Parkville

Melbourne, 3010

Australia

Phone: 61 405761879

Email: sedigh.khademi@gmail.com

Abstract

Background: Collecting information on adverse events following immunization from as many sources as possible is critical for promptly identifying potential safety concerns and taking appropriate actions. Febrile convulsions are recognized as an important potential reaction to vaccination in children aged <6 years.

Objective: The primary aim of this study was to evaluate the performance of natural language processing techniques and machine learning (ML) models for the rapid detection of febrile convulsion presentations in emergency departments (EDs), especially with respect to the minimum training data requirements to obtain optimum model performance. In addition, we examined the deployment requirements for a ML model to perform real-time monitoring of ED triage notes.

Methods: We developed a pattern matching approach as a baseline and evaluated ML models for the classification of febrile convulsions in ED triage notes to determine both their training requirements and their effectiveness in detecting febrile convulsions. We measured their performance during training and then compared the deployed models' result on new incoming ED data.

Results: Although the best standard neural networks had acceptable performance and were low-resource models, transformer-based models outperformed them substantially, justifying their ongoing deployment.

Conclusions: Using natural language processing, particularly with the use of large language models, offers significant advantages in syndromic surveillance. Large language models make highly effective classifiers, and their text generation capacity can be used to enhance the quality and diversity of training data.

(JMIR AI 2024;3:e54449) doi:[10.2196/54449](https://doi.org/10.2196/54449)

KEYWORDS

vaccine safety; immunization; febrile convulsion; syndromic surveillance; emergency department; natural language processing

Introduction

Background

A febrile convulsion refers to a seizure triggered by a fever, most commonly experienced by children aged between 6 months and 5 years, in the absence of an underlying central nervous system infection or metabolic disturbance [1]. Febrile convulsions have various causes and risk factors, including viral or bacterial infections, a family history of seizures, underlying neurological conditions, environmental factors, and specific vaccinations [2]. Age, fever, and a seizure are essential components of the definition of childhood febrile convulsion [3]. Febrile convulsions are most often caused by viral respiratory tract infections but are also associated with viral infections such as chicken pox, tonsillitis, and middle ear infections. Febrile convulsions are also associated with the administration of childhood vaccines [4]. Although febrile convulsions caused by vaccines are rare and typically do not cause permanent damage, parents' experiences with their children's febrile convulsions can have a negative effect on their perception of vaccine safety [5].

In 2010, in Australia, there was an increase of febrile convulsions in young children after the release of the southern hemisphere trivalent inactivated influenza vaccine, produced by CSL Biotherapies [6,7]. Following national suspension of seasonal influenza vaccinations for children aged <5 years, reviews [8,9] revealed deficiencies in Australian adverse event following immunization (AEFI) monitoring system, which had resulted in delayed reporting and underreporting of febrile convulsions [10,11]. The reviews highlighted the need for monitoring additional data sources for early AEFI detection, a subsequent focus of Surveillance of Adverse Events Following Vaccination In the Community [12,13] and the Health Informatics groups at the Murdoch Children's Research Institute, Victoria, Australia. A recent paper highlighted the need for vaccine safety monitoring to include natural language processing (NLP) of both internet-based data sources and electronic health records [14].

In this study, we aimed to assess the effectiveness of NLP techniques for rapid detection of febrile convulsion presentations in emergency departments (EDs).

Syndromic surveillance relies on the categorization of patient-presented symptoms and complaints into "syndromic indicators," often derived from patient-reported or observed symptoms [15]. These indicators, recorded by health care providers during the initial patient contact, along with preliminary or working diagnoses, are crucial in the absence of any confirmatory testing or diagnosis to facilitate prompt public health decisions [16]. Examples include monitoring of telephone health advice systems, of notes taken during attendance to primary physicians, and of data entry performed during visits to ED.

Syndromic surveillance has shown to have the ability to rapidly evaluate the potential impact of a recently introduced vaccine [17,18]. Monitoring telephone helpline data can also assist with early detection of AEFI, and in the case of 2010 Australian

AEFI signal, retrospective analysis of these data showed that such methods would have flagged a signal 2 weeks after commencement of vaccination, which is 4 weeks earlier than the alert was raised [19]. Surveillance of ED triage notes is particularly effective for timely syndromic information capture, as data are entered upon a patient's arrival to ED, allowing for the initiation of a notification from a surveillance system while the patient is still in the ED [20], well before any diagnostic coding takes place.

ED triage notes are gathered during the first moments of the patient encounter and usually contain aspects of a patient's medical history, presenting symptoms, and the reasons for their visit. This information is primarily used to direct initial clinical management and can serve as a tool to help understand trends of patient visits in near real time [21]. However, variation in the language used in the documentation of this information within and across hospitals significantly impedes the reuse of these data [22]. Abbreviations abound and their meanings vary according to context; for example, "cp" might be used as abbreviation for any type of *chest* pain, which can include pulmonary and trauma-related sources, in some contexts, it might just mean *cardiac* pain, while in others, it may refer to cerebral palsy. In some presentations, "NVD" means "nausea, vomiting, and diarrhea," but in relation to childbirth, "NVD" means "normal vaginal delivery." Misspellings, local variations of abbreviations, and context-sensitive vocabulary all feature in ED notes, and there are additional variations of the quality and length of the texts [23].

Research examining triage notes can be broadly classified into 3 main categories: quality improvement of triage notes' recording and category assignment, prediction, and case identification. Studies focusing on the quality improvement of triage notes' recording and category assignment aim to enhance the accuracy, reliability, efficiency, and completeness of the information recorded during triage. Prediction studies aim to predict the outcomes of emergency visits or the resources needed by patients based on the information recorded in triage notes. Case identification studies aim to either classify ED visits into categories or syndromes or to collect data about specific presentations (or syndromes) of interest [24].

Various methods have been used for identifying syndromes from triage notes. These include keyword-based, linguistic-based, statistical and machine learning (ML) algorithms, or hybrids of these [25]. Some have used data from 1 hospital [26], while others have used data from >1 hospital [27]. These systems vary in their goals, such as focusing on classification of 1 syndrome [28] or developing a syndromic surveillance system for >1 syndrome [29].

In recent years, the use of ML algorithms for surveillance of ED triage notes has increased [24]. One of the main obstacles in using supervised ML algorithms is the scarcity of annotated data for training and benchmarking [30]. Many studies have used medical coding against existing data as a proxy for the labels [31-34]. However, the use of *International Classification of Diseases* codes as a gold standard has known limitations as they do not always align with the actual reason for the visit [35]. For instance, codes can be assigned to identify the underlying

etiology of a presentation, for health conditions not directly observed in the text of a presentation, or for other purposes such as financial incentives [36,37]. The choice of codes can be influenced by perceptions of the importance of a certain condition [38], and in this study, a febrile convulsion might not get coded if it is thought of as a secondary effect or not significant enough to include on a discharge summary when there are limitations to how many codes may be assigned.

When using supervised ML algorithms in the context of syndromic surveillance of ED triage notes, researchers have manually annotated from several thousand [29] to a few hundred thousand records [39] to train the algorithms. It has been observed that data annotation poses a significant obstacle in training NLP models within the clinical domain, with manual identification of labels affecting the representativeness of samples. This challenge often restricts NLP solutions to obtain data from only a few institutions, thereby impacting their generalizability [40].

Objectives

In this study, the overarching aim was to identify emerging trends that could signal potential issues with a vaccine. Our primary objective was to construct a highly effective NLP model for the early detection of febrile convulsions in ED notes, applicable to the entirety of public hospital ED departments, without requiring large volumes of annotated training data. We achieved this goal by leveraging a limited set of manually labeled records and using data augmentation techniques. Our data set was sourced from 26 public hospitals across the Australian State of Victoria. Furthermore, we aimed to outline the essential requirements for the development and deployment of such a system.

Methods

Data

Overview

SynSurv provided the primary data source for this study. SynSurv is the syndromic surveillance project of the Department of Health of the state government of Victoria, Australia. Its objective is to detect events of public health significance early, allowing clients responsible for public health action to respond promptly and effectively. At the time of writing, SynSurv receives a rapid stream of information about every ED presentation, including the triage text, from a majority (n=34) of the public hospitals with Emergency Departments in Victoria, Australia. Most presentations arrive within 5 to 15 minutes of the patient's assessment.

Data comprise the text recorded at triage by ED nurses and are characterized by a unique structure that primarily consist of abbreviations and brief phrases. The text usually contains a presenting complaint, selected past medical history, and the nurse's observations of the patient. Triage text does not contain demographic or identifying information. The length of the text varies; it may be a detailed narrative of the patient's presentation to the triage nurse or it could be a concise summary of a possible diagnosis along with a few observations. The unlabeled data

set used in this study consisted of 76,274 ED triage text from January 1 to July 14, 2022, of ED presentations of children aged between 6 months and 6 years. The average length of text was 22 (SD 20.4) words. The longest record initially contained 319 words, but after data preparation, which involved removing nontextual information, the length of the longest text was reduced to 253 words. Additional data collected in 2022 were used to create a hold-out test data set.

Febrile Convulsion Symptoms

During the initial, "tonic" seizure stage of a febrile convulsion, the individual may let out a cry or moan before suddenly losing consciousness and experiencing muscular rigidity. This stage can last for up to 30 seconds and may be accompanied by the cessation of respiratory movements. The "clonic" seizure stage that follows involves repetitive movements of the limbs or face. While rigors (uncontrolled shivering and shaking) may look similar and often occur during any acute febrile illness, loss of consciousness is not typically associated with them [41]. Seizures typically last <5 minutes, although they may be prolonged. A "postictal state" follows, lasting between 5 and 30 minutes, during which the patient can experience drowsiness, confusion, headaches, and nausea while gradually returning to normal.

Data Annotation

Annotation of febrile convulsions needs to account for the language used in their clinical descriptions, which includes temperature-related terms and terms used to describe the clonic, tonic, and postictal stages of a seizure.

The first step of annotation involved filtering the ED notes for convulsion-related terms (eg, "seiz," "convuls," "fit," "epilep," "ictal," "tonic," or "clonic") and fever-related terms (eg, "febri," "fever," "37.," "38.," "39.," "40.," "41.," "42.," or "43.>"). Applying the filter reduced the data set to around 29,000 candidates for labeling, and these were annotated with a goal to identify around 1000 positive examples of febrile convulsion. The ED nurse's notes were thoroughly reviewed, and if there was a likely indication of a febrile convulsion, whether explicitly mentioned or not, a positive label was assigned by J Black (described below), and a negative label was assigned to records that did not meet the criteria. In an additional step, some records that did not contain any of the filter settings were randomly selected and labeled. Only a few of these were identified as positive, mostly due to spelling variations in the filter strings that caused the records not to be detected in the initial step.

An annotation guideline was developed by J Black, who is a physician with ED experience, where a record was labeled as positive if the following criteria were met:

- The patient presented with febrile convulsion symptoms at the time of ED presentation, which requires mentions of both seizure and fever.
- The mention of febrile convulsion is not just in the patient's medical history (eg, only "phx febrile convulsion") or just an expression of parental concern (eg, only "mother worried as child previously had a seizure").
- The convulsion is not related to other chronic conditions that include seizures, as febrile illness can trigger a

pre-existing disposition to seizures. A mention of medicine usually taken when seizures happen is an indication of existing underlying cause.

- A mention of fever-lowering medications, subjective assessment of fever by parents or carers, or measurements taken at home can indicate the presence of a fever, even if the temperature recorded in the ED is normal.
- The notes do not indicate other types of seizure, including absent or focal seizures.

Following this guideline, author J Black annotated the training data to classify instances as either febrile convulsion or not. This resulted in 1032 positive labels and 14,415 negative labels, making a total of 15,447 annotations. The annotation of the separate test data set resulted in 432 positive and 2768 negative labels, a total of 3200 records. [Table 1](#) provides examples of triage notes along with their corresponding labels, and [Table 2](#) enumerates the record and word counts of the data sets.

Table 1. Sample of triage notes (not the actual text but examples of typical structure).

Label	Category	Triage notes
1	Fever and seizure present	“Seizure, Unwell Since yesterday, Febrile, Vomit enroute, IUTD, Nil Rash, A-PATENT, Nil Sob, T- 38.8, GCS-15, R- 22, Nil Pain, PWD, O/A alert.”
1	“Tonic clonic” and fever but no mention of convulsion or seizure	“BIBA: Unwell 1/12 (fev,diahroea). 1× ep of eye rolling back? tonic clon. Self res 1/60. Good oral intake. O/E: PWD, Good capp Asleep, easily rouse. Ket 0.4 Pmhx: UTDI”
1	Fever and spelling variation of seizure-related term and fever	“FEVER. Decreased oral intake. Sz today. Runny nose. Clear chest. Nil vomiting. Miserable at triage. Phx dad states rare gene mutation.”
1	Seizure and spelling variation of fever-related term	“BIBA post seizure tonight, eyes rolled and floppy unresponsive 5 min. Temps last 3/52. At triage febrile.”
1	Seizure and fever medication is mentioned	“At midnight Advil given. At 0100 10 sec of seizure like activity. 2nd seizure activity shortly after lasting a few minutes with bilious vomit. O/A alert. RR 24 sats 99 no WOB HR 112.”
0	Underlying condition for seizure	“Prolonged sz at home lasting 13 mins. Congested left lung with resp symptoms. Runny nose and cough this week. 4mg oral midaz. Resolved with av arrival. Postictal 40 mins. Phx sz focal or tonic/clonic phx eplispey and dravat sx”
0	Febrile convulsion in past medical history only	“BIBA seizure like activity tonight, intermittent 15 mins, post ictal .30/60. Hx febrile convulsions. Afebrile oa”
0	Parental concerns	“FEVER Since yesterday, Coughing, Nasal congestion, Parents concerned as previous febrile convulsion”
0	Other types of seizures	“FEVER? Absent episode at day care following fever. Nil seizure activity, but more quiet. GCS 15. pHx asthma”

Table 2. Descriptive statistics of the data set.

Data set	Total records, n	Total words, n	Average words, n	Unique words, n
Initial data	76,274	1,654,045	21.69	52,698
Training data	15,447	398,608	25.80	21,086
Test data	3200	102,625	32.07	9925

Data Set Construction

The original data were very imbalanced, with 1032 positive labels versus 14,415 negative labels. Our approach was to allow for the influence of the negative examples as much as the positive examples, as we wanted to ensure the models were not overly prone to identify false positives. We decided to evaluate as many negatively labeled records as possible by dividing the negative data into smaller data sets that each roughly matched the number of available positive records, while also matching the positive records’ text lengths, and we paired each of these with the positive records for a training data set. To accomplish text length matching, we assigned a word-count group to each record, and when sampling negative records, we ensured that we took similar numbers from each word-count group to those which existed in the positive records.

After setting aside a validation data set of 100 positive and 100 negative labels, 932 positive labels remained for training. When sampling the negative labels to match the 932 positive labels for a training data set, we oversampled by a factor of around 1.2, iteratively extracting negative examples until we ran out of examples to complete a training data set. The result was 9 data sets each consisting of the same 932 positive labels and 1127 different negative labels, with a total of 2059 records in each. These were used for initial training of 9 identical transformer models, and by assessing their test scores, we could determine that the best model also identified the best of the 9 training data sets, where the balance of negative examples worked most advantageously with the positive examples. We chose transformers for the training data evaluation because with their capacity to take account of language structure, they were more sensitive to textual information compared to the other standard neural networks we were evaluating [42].

Data Augmentation

When evaluating the transformer models, we found that there was potential for improvement in their performance, as their F_1 -score was around 0.79. Therefore, we decided to assess the effect of training with additional examples. As we lacked new positive examples, we decided to experiment with data augmentation techniques. These included synthetic text generation using GPT-2 models, domain-specific data augmentation, and task-agnostic techniques. This is explained in a prior publication [43] where the best result was achieved by using synthetic text generation techniques. Using this approach added 1582 positive labeled records to the training data set. This meant we could safely add further negative examples without the data set becoming imbalanced, resulting

in 5455 training records. We used this augmented data set and the 200-record validation data set to train and validate all the approaches, and we evaluated and compared them using the 3200-record holdout test set.

After conducting error analysis on the predictions of the transformer model, we added another 112 seizure-related negative labeled records to the data set. This gave the models additional exposure to ED notes about other types of seizures or tonic-clonic seizures without the fever components, which would allow to the model to better learn not to create false positive predictions on these marginally negative examples. The final training data set consisted of 2514 positive and 3053 negative labels, a total of 5567 records. Table 3 shows the construction of the data sets.

Table 3. Data sets' construction.

Data set	Positive, n (%)	Negative, n (%)
Initial training ($\times 9$; n=2059)	932 (45.3)	1127 (54.7)
Augmented training (n=5455)	2514 (46.11)	2941 (53.89)
Final training (n=5567)	2514 (45.22)	3053 (54.78)
Validation (n=200)	100 (50)	100 (50)
Test (n=3200)	432 (13.5)	2768 (86.5)

Classification

Overview

As part of our goal of determining the most effective NLP methods for identifying febrile convulsions in ED visits, we needed to assess the trade-off in requirements and benefits of increasingly sophisticated NLP models. While the Data section described the data requirements that evolved as these models were assessed, the Classification section describes the reasons for evaluating the various models, their data preparation and training requirements, and some relevant observations of their training processes. Evaluation and results will be discussed fully in the Results section.

Pattern Matching

We started with pattern matching as a manual approach that could give us a baseline against which we could compare ML approaches. This consisted of selecting text based on relevant strings that would, when combined, indicate both fever and convulsions. After importing the data into a Structured Query Language (SQL) database, we used a SQL full-text search to describe the patterns more comprehensively and to ensure that both fever and convulsions were included together. The pattern used was ('fever*' OR 'pyrex*' OR 'feb*' OR '37.6*' OR '37.7*' OR '37.8*' OR '37.9*' OR '38.*' OR '39.*' OR '40.*' OR '41.*' OR '42.*' OR '43.*' OR 'T38*' OR 'Temp38*' OR 'T39*' OR 'Temp39*' OR 'T40*' OR 'Temp40*' OR 'T41*' OR 'Temp41*' OR 'T42*' OR 'Temp42*' OR 'T43*' OR 'Temp43*' OR 'hot*' OR 'warm*') AND ('convul*' OR '*seiz*' OR '*size*' OR '*sezi*' OR ictal OR tonic OR clonic). We experimented with accounting for negations and modifiers, such as mentions of medical history and temperature measurement units, when they were close to terms related to

febrile convulsions. Although this improved the detection of false positives, it detrimentally affected the detection of true positives and ultimately resulted in a poorer performance of the model. Therefore, to retain the simplest and most performant pattern matching approach, we decided to avoid dealing with negations and modifiers.

Standard Classifiers

We evaluated a variety of standard classifiers using both the original text with trigrams and lemmatized forms of trigrams. Given the highly specific nature of the texts, where abbreviations and punctuations prevail, we did not eliminate stop words, and we restricted our preprocessing to (1) expanding any contractions that used apostrophes and (2) converting collections of the plus sign (eg, "+++") into the word "extreme," as these are used throughout to convey that meaning. We used a customized tokenization method, based on spaCy, to ensure that tokenization did not break apart symbols on embedded periods and slashes and to fashion bigrams, trigrams, and lemmatized words. This was because we needed to preserve the forms of words, especially regarding temperatures and explicit compound expressions, and to control how n-grams and lemmatization were performed. We experimented with the Scikit-Learn CountVectorizer and TfidfVectorizer, with and without inverse document frequency (IDF) enabled; with the result that for each of the 3 data sets, we had standard, trigram, and lemmatized versions, each of which were assessed via a CountVectorizer, TfidfVectorizer with IDF, and TfidfVectorizer without IDF enabled; a total of 27 data sets for each of the classifiers were assessed.

Standard Neural Networks

We preprocessed the data for experimentation with standard neural networks by separating basic contractions (eg, isn't to is

n't), but otherwise, the data were left intact, with no ED-specific translations done per the Standard Classifiers section, such as “+++” to “extreme.” The preprocessed text was tokenized using the *torchtext* library, and models were constructed using the Pytorch library. We experimented with a range of neural network models—convolutional neural network (CNN), long short-term memory (LSTM), bidirectional LSTM (BiLSTM), CNN-LSTM, and CNN-BiLSTM hybrids. Then, we tested the gated recurrent unit (GRU), bidirectional GRU (BiGRU), and CNN-GRU and CNN-biGRU hybrids. We implemented the best of these—the BiGRU and the CNN-BiGRU hybrid. The latter consisted of a 3-layer CNN producing one-grams, bigrams, and trigrams (via kernel sizes of 1, 2, and 3 with a kernel number of 100), with its output concatenated with a BiGRU output. The standard BiGRU model's F_1 -score slightly exceeded that of the CNN-BiGRU hybrid model, but the CNN-BiGRU was included because it mostly had a better balance of precision and recall and was a strong contender.

Transformers

We used the RoBERTa-large-PM-M3-Voc model, published by Facebook [44] and described as being “pre-trained on PubMed and PMC and MIMIC-III with a BPE Vocab learnt from PubMed.” This model was selected due to its superior performance in classifying biomedical and clinical texts compared to other models with similar capabilities, including Scientific Bidirectional Encoder Representations from Transformers (SciBERT) by the Allen Institute for Artificial Intelligence, Biomedical (BioBERT) by researchers at Korea University and the National Institutes of Health, ClinicalBERT by researchers at the University of Pennsylvania and the University of Washington, and BioMed-RoBERTa by researchers at the University of California, San Diego. We did no text preprocessing, as we considered that the transformer's internal byte pair encoding approach and inherent language understanding as sufficient to deal with the texts' complexity. The best transformer model was identified from the final form of the training data.

Ethical Considerations

Ethics approval for this study was granted by the Department of Health, Human Research Ethics Committee in Victoria, Australia (project ID: HREC/83486/DOH-2022-298485). No compensation was provided to any participants. Informed consent was not sought for this study because the operational work it supports aligns with legislation related to serious public

health threats. The data were anonymized by removing personal details and using a 1-way hashing algorithm to ensure that reidentification is not possible.

Results

Overview

The results are shown in Table 4 as precision, recall, and F_1 -scores when evaluated against the test data set. We used precision, recall, and F_1 -scores as evaluation metrics to assess the performance of the models on the positive label using the test data set. Precision measures the proportion of correctly classified positive instances out of all instances predicted as positive. Recall measures the proportion of correctly classified positive instances out of all actual positive instances. F_1 -score is the harmonic mean of precision and recall, providing a balanced measure of performance. By using these evaluation metrics, we were able to comprehensively evaluate the models' performance on key measures of accuracy and completeness with respect to both the positive and negative labels.

Scores are depicted for each model in order of the data sets used to train a model. These were (1) the best of the initial training data set of 2059 records, (2) the synthetic records-enhanced data set of 5455 records, and (3) the data set also containing 112 additional examples of negative seizure examples, with a total of 5567 records. These data sets are indicated with superscripts of b, d, and e in Table 4.

The models' difference scores are shown at the bottom of each model group. These are calculated as the difference between model test scores obtained when trained on the final (ie, third) data set (superscript e) and the test scores obtained using the pattern matching approach, which functions as a baseline. The best individual value in each of the table columns are in italics, but the F_1 -score is the most important value to measure overall performance.

Figure 1 shows a graphical comparison of the F_1 -scores achieved per model on the test data set, as the models were trained on the 3 training data sets. Notably, the F_1 -score of the RoBERTa transformer model was initially no better than pattern matching when trained on the first data set. However, as more data were added, the RoBERTa transformer model's performance improved significantly, surpassing the F_1 -scores of the other models starting from the second data set onward.

Table 4. Model performance metrics.

Model and data sets	Precision	Recall	F_1 -score	True negative	True positive	False positive	False negative
Pattern matching	0.665	<i>0.993^a</i>	0.797	2552	429	216	3
RoBERTa ^{b,c}	0.658	<i>0.993</i>	0.792	2545	429	223	3
RoBERTa ^d	0.852	0.972	0.908	2695	420	73	12
RoBERTa ^e	0.875	0.972	<i>0.921</i>	2708	420	60	12
RoBERTa—difference ^e	0.210	-0.021	0.124	156	-9	-156	9
BiGRU ^{b,f}	0.812	0.917	0.861	2675	397	92	36
BiGRU ^d	0.866	0.921	0.893	2705	399	62	34
BiGRU ^e	<i>0.903</i>	0.898	0.900	2725	389	42	44
BiGRU—difference ^e	0.237	-0.095	0.104	173	-40	-174	41
CNN-BiGRU ^{b,g}	0.822	0.915	0.866	2681	396	86	37
CNN-BiGRU ^d	0.889	0.887	0.888	2719	384	48	49
CNN-BiGRU ^e	0.864	0.937	0.899	2742	374	59	25
CNN-BiGRU—difference ^e	0.199	-0.056	0.102	190	-55	-157	22
XGBoost ^{b,h}	0.633	0.940	0.757	2533	406	235	26
XGBoost ^d	0.723	0.928	0.813	2614	401	154	31
XGBoost ^e	0.746	0.917	0.822	2633	396	135	36
XGBoost—difference ^e	0.081	-0.076	0.026	81	-33	-81	33

^aItalicized values represent the best individual value in each of the columns.

^bThe best of the initial training data set of 2059 records.

^cRoBERTa: Robustly optimized Bidirectional Encoder Representations from Transformers approach.

^dThe synthetic records-enhanced data set of 5455 records.

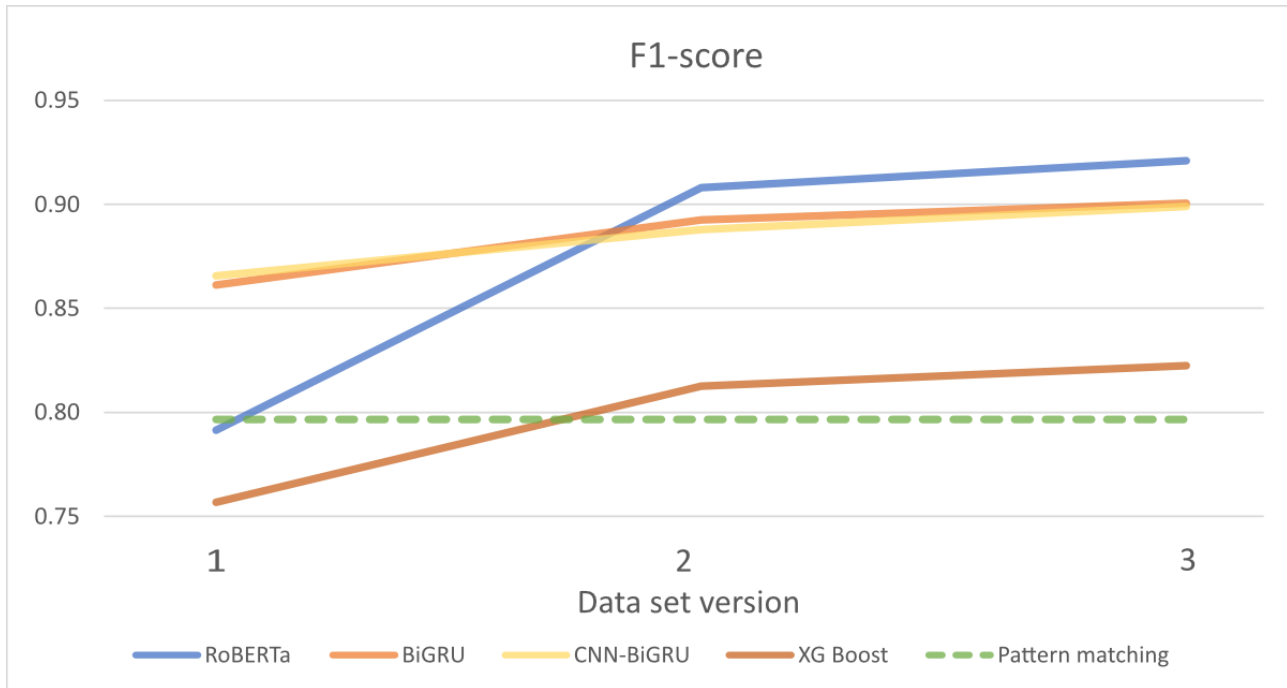
^eThe data set also containing 112 additional examples of negative seizure examples.

^fBiGRU: bidirectional gated recurrent unit.

^gCNN-BiGRU: convolutional neural network-bidirectional gated recurrent unit.

^hXGBoost: extreme gradient boosting.

Figure 1. Comparison of the F_1 -scores of the models. BiGRU: bidirectional gated recurrent unit; CNN-BiGRU: convolutional neural network-bidirectional gated recurrent unit; RoBERTa: Robustly optimized Bidirectional Encoder Representations from Transformers approach; XGBoost: extreme gradient boosting.



Pattern Matching

The pattern matching method we assessed was a rules-based approach looking at text patterns in a SQL full-text query. It achieved a very high recall of 0.99—meaning it correctly identified almost all the febrile convulsion records. However, it also identified many incorrect records as febrile convulsions with a resulting precision of 0.67, so its F_1 -score suffered, although it remained acceptable at 0.80 (rounded). Analysis of false positives showed that mentions of time of the day, duration of seizure, and child weight were all open to misinterpretation as indications of fever, as numbers ranging from “37.6” through to “43.” were matched as indicators of temperature. By contrast, references to temperature that had no accompanying qualifier or decimal point were ignored. Terms such as “warm to touch” were missed, where warm can refer to fever (but is also used in observations about skin being pink, warm, and dry). Imprecise descriptions of febrile convulsion, negations, and mentions of previous history of fever, seizure, and febrile convulsions were missed or misinterpreted. For instance, implied and specific references to history such as “increased seizure,” “mother concerned,” “Mum states older brother has had a febrile seizure before,” “febrile convulsion previously,” “recent admission with seizure + enterovirus,” and mentions of use of seizure-related medications before the emergency visit. We experimented with including pattern matching of previous history in our query, which worked to a certain extent to remove false positives but also removed true positives, with a resulting worse performance. A fully implemented pattern matching system requires extensive rules that adjust for many textual nuances but which can never be complete and would be difficult to maintain, which is why we focused our effort on ML solutions.

Standard Classifiers

We assessed the Scikit-Learn Multinomial Naive Bayes, logistic regression cross validation, linear support vector classification (SVC), stochastic gradient descent (SGD), random forest, extra trees classifiers, and the extreme gradient boosting (XGBoost) classifier. Each was tested with the 3 data sets; with the standard form of the text, with trigrams, and with lemmatized trigrams; and vectorizing with the Scikit-Learn CountVectorizer, TfidfVectorizer with IDF, and TfidfVectorizer without IDF enabled. Grid searches were performed to further tune model parameters for the best models from each round. We evaluated the models with the test data set.

On the initial data set of 2059 records, the best model was the XG Boost classifier, using lemmatized text and the CountVectorizer, with an F_1 -score of 0.757. The second best was the logistic regression cross validation model, using standard text and the CountVectorizer, with an F_1 -score of 0.755. The XG Boost model continued to be the best model as we assessed with the larger data sets. With the augmented, second data set of 5455 records, which included synthetic positive examples, the F_1 -score was 0.813, using standard text and the TfidfVectorizer with no IDF. With the third data set of 5567 records, which included extra negative seizure examples, the F_1 -score was 0.822, again using standard text and the TfidfVectorizer but this time with IDF enabled.

The standard classifiers were worse than pattern matching on the initial data set, but on the larger data sets, the best of the models scored better than pattern matching. The XG Boost model fared well for *recall*, with an average of 0.93 over the data sets, which was marginally better than the standard neural networks’ average of 0.91 but not as good as the transformer’s average of 0.98. However, the superior precision of the standard

neural networks resulted in their F_1 -scores being, on average, 0.09 higher than those of the XGBoost model, while the average F_1 -score of the transformer model was 0.10 higher. With a relatively good recall but overall poorer performance and a high degree of effort required to prepare for and assess these models, we would consider these only in a low-resource situation and would consider pattern matching as a comparable option.

BiGRU and CNN-BiGRU

The BiGRU and CNN-BiGRU classifiers were Pytorch models trained from scratch on the different data sets. Other models that were evaluated did not come close to their performance; these included a CNN model, LSTM models (in both standard and bidirectional form), CNN-LSTM, CNN-BiLSTM, and a BiGRU model with an additional attention layer. Word2Vec embeddings from the training data sets were loaded into the models.

Their performance based on F_1 -score was as much as 10% better than that of pattern matching at 0.8, ranging from around 0.86 to around 0.9 as the data sets were developed. This was because their precision was much better; they found fewer false positives, although their recall was poorer compared to pattern matching by as much as 10%.

Initially, they were better than the transformer approach as well; their starting F_1 -score of around 0.86 exceeded the transformer's score of 0.79. However, as data were added, first through additional synthetic positive records and then by adding negative seizure examples, these models only improved over their initial scores by around 4 percentage points (lower than the improvement of transformer, discussed next).

Transformers

The transformer model architecture has been proven to be very suitable for fine tuning to tackle language tasks, with previous research [45] demonstrating that these models outperform other neural networks and standard classifiers. The RoBERTa-large-PM-M3-Voc model was chosen because of its proven capacity to understand clinical texts. Our initial training used 9 slightly imbalanced data sets, all with the same 932 positive labels but each with different 1127 negative labels, which, at 2059 records, was scarcely enough to fine-tune a transformer. However, we were able to establish which of the data sets worked best with the model, and this data set then became the foundation of data set development, which was chiefly undertaken to improve the performance of the transformers while providing a comparison to their performance improvements against other models.

The initial transformer model did no better than the pattern matching; its F_1 -score was 0.792, which was slightly less than the 0.797 of the pattern matching and much less than the 0.866 of the CNN-BiGRU model. Encouragingly, it matched the pattern matching model's recall of 0.933, but it was let down by a relatively poor precision of 0.658. We found that this was mainly due to false positives; it was classifying many seizure-related records as febrile convulsion when they had no mention of fever. We added 112 more examples of negatively labeled seizure records, which we had manually checked to

ensure fever was absent. The expectation was that the model could learn that seizure alone was not predictive of a febrile convulsion. However, this had only a slight effect; the F_1 -score only increased by 0.6 percentage points to 0.797, now equaling the pattern matching.

Therefore, we decided instead to add a lot more positive and negative records, as the model was clearly struggling with a lack of data to learn from. As we had no more positive records, we used a GPT-2 language model to generate synthetic examples of positive labels, as described previously [43]. Adding these enabled us to also sample and add many more negative labels to get a reasonably balanced data set of 5455 records, containing 2514 positive and 2941 negative labels—the slight imbalance was to give the model more negative examples. Training the transformer and other models just on the combined initial and synthetic data allowed us to measure the effect of the synthetic records clearly. Any extra negative records were randomly selected and not taken from the manually crafted 112 seizure negative records, which we had set aside at this point. This had a very positive effect; the F_1 -score on a newly fine-tuned model increased by 11.7 percentage points to 0.908, beating the 0.893 of the BiGRU, which was the best performing standard neural network trained on this version of the training data set.

Finally, we readded the 112 negative seizure records and reran our training for all models; the best transformer model now achieved an F_1 -score of 0.921, a more significant 1.6 percentage points improvement compared to the 0.6 increase obtained when the negative seizure records alone had been added in our previous experiment with them.

The best performing transformer model still maintained an impressive recall of 0.972, compared to its initial 0.993, but its precision had risen to 0.875 from 0.665. Its F_1 -score of 0.921 was 12.4 percentage points better than the pattern matching baseline of 0.797 and 2 percentage points better than the best standard neural network, which was the BiGRU with a score of 0.900.

This result confirmed our previous experience that transformers need more data to learn from compared to lighter-weight neural networks, but in the right conditions will outperform those simpler architectures. While the neural networks improved over their initial scores by around 4 percentage points following addition of synthetic data, the transformer improved from its starting score by >3 times that amount, at around 13 percentage points. The final transformer was superior to the final BiGRU by 2 percentage points (0.92 vs 0.90).

Deployment

The best performing models from both standard neural networks and transformers were deployed in a Databricks environment. The transformer model was originally 1.4 GB in size and ran in a timely manner best on a graphics processing unit (GPU); inference was 5 times faster on a GPU versus a central processing unit (CPU). For performing inference on a CPU, we wanted to reduce both the memory required and the speed to perform inference. Therefore, we converted the best transformer checkpoint to an Open Neural Network Exchange (ONNX) model and optimized and then quantized it. Although the

optimized model was no smaller than the original transformer model, it ran twice as fast as the transformer on a CPU. The quantized model was considerably smaller at 500 MB, had no loss of accuracy, and ran 40% faster again than the optimized model on a CPU; hence, it was used. Inference times on CPU for the quantized ONNX model were similar to the transformer model on GPU. However, inference times on CPU using the CNN-BiGRU, which was only 14 MB in size, were 10 to 12 times faster than the transformer on GPU and the quantized ONNX models.

The minimum available Databricks configuration was a Standard DS3 v2 CPU compute with 14 GB of memory and 4 cores and a 13.1 ML runtime, which costed 0.75 Databricks units per hour. All models were able to run on this. However, the much faster inference and smaller memory requirement of the standard neural network models meant that this configuration would be able to support the parallel loading of many such models (for the surveillance of numbers of syndromes), while based on the use of 2 transformer-based ONNX models, we estimate that it would support only up to 6 simultaneously loaded models before requiring a parallel deployment of computing capacity or a more powerful single capacity.

After 5 months of deployment, the model has predicted 749 febrile convulsion cases in the target group of children aged <6 years and 75,543 cases of no febrile convulsion in that group. To evaluate its performance, we sampled 125 of its predictions for each cohort. To ensure we had good candidates for potential false negatives, when sampling for the 125 nonfebrile convulsions, we filtered to records that had a mention of either febrile, seizure, or convulsion. Labeling resulted in 122 febrile convulsion and 128 nonfebrile convulsion records. The model had predicted incorrectly for 9 (3.6%) of the 250 records, resulting in a precision score of 0.952, a recall score of 0.975, and an F_1 -score of 0.964.

Discussion

Overview

The key objective of this study was to contribute to improving near real-time syndromic surveillance of febrile convulsions by using NLP models. We compared NLP approaches with a pattern matching baseline solution. We found that even with minimal initial training data but careful attention to the training examples and the addition of augmented data to improve the data, a transformer-based model could achieve superior performance, without needing any demanding text preprocessing or feature construction. We concluded that while the process of determining the best training data set was nontrivial, the result justified the effort and acted as a guide to further development for these models for classifying ED notes.

Principal Findings

The format, quality, and length of ED triage notes can differ greatly, which presents a considerable challenge when it comes to text processing. To overcome these dissimilarities, one solution is to use lexicons to replace variations in spelling, abbreviations, and medical terms with standardized synonyms, and another solution is to use rules to recognize specific text

features. Nonetheless, both these methods demand ongoing efforts to handle novel words or establish new rules, which can make the system more intricate as additional rules often need to be introduced to amend the impact of previously applied rules. In addition, the use of these methods can negatively affect the generalizability of solutions across hospitals, as terminology and abbreviations can be specific to individual ED departments.

In our research, we have demonstrated that neural networks and especially cutting-edge large language models can remove the need for preprocessing of text and can use text as is to achieve outstanding performance in syndromic surveillance of ED notes. Large language models have been originally trained on substantial volumes of text and have extensively learned complex textual patterns and relationships within texts, and when fine-tuned, they can quickly learn the specifics of previously unseen texts such as triage notes. This learning is enhanced if the model has been pretrained on similar texts, as was the case with the RoBERTa-large-PM-M3-Voc model we used, which had been trained on biomedical and clinical texts.

Development of supervised algorithms requires labeled data, which is hard to acquire [40,46]. Various techniques have been used by researchers to overcome this barrier. Researchers have employed various techniques to overcome this barrier, ranging from using proximal ICD codes [31], which suffers from a loss of expert targets, to the labor-intensive process of manually labeling many records [39]. Our strategy showed that use of language models for generating synthetic text is a highly effective and efficient way to augment data to improve the performance of the classification task. However, our findings suggested that although data augmentation can have a significant impact on the performance of language model-based classifiers, its impact on more conventional classifiers such as CNNs may be more limited. Specifically, augmenting data can substantially improve the accuracy and robustness of language model-based classifiers by expanding the data set and introducing greater variation in the data, particularly when there is information in the texts that clarifies the features of classes. However, there are lesser gains realized by standard neural networks and traditional classifiers, which indicates the greater ability of language models to benefit from textual clues.

Our findings also suggested that there is no single solution that can be universally applied for syndromic surveillance of ED triage notes, and simple pattern matching may provide reasonable performance, particularly where a syndrome can be clearly identified with the presence of specific keywords.

Clinical NLP research has been ongoing for several decades and has contributed significantly to many areas of patient care. However, despite these advances, there is still a lack of NLP systems that have been deployed and integrated into operational settings [47]. Our solution is currently deployed in a cloud-based environment and is continuously sending a stream of flagged presentations to an organization tracking adverse events following immunization monitor for possible increases against their background rate for detection of any vaccine safety signal related to febrile convulsions.

Limitations

Our approach is extendible to similar scenarios; however, the model we created is specific to the task of detection of febrile convulsions in ED notes. Although the process of careful analysis, leading to an informed application of methods to enhance training data, is repeatable for the detection of other syndromes and potentially beyond (eg, vaccine adverse events following immunization), the approach depended on personal judgment and experience and was somewhat complex. More methodical approaches to determining optimal training data are described in the active learning literature [48], and our future

focus will be on implementing these approaches while leveraging the insights gained from this study on using augmentation to enhance training data.

Conclusions

Near real-time surveillance of febrile convulsion presentations to EDs is feasible using NLP solutions. We established that a large language model classifier can be trained in the context of few training examples by adding synthetically generated data and implemented into a real syndromic surveillance system, enabling surveillance of febrile convulsion following vaccination.

Acknowledgments

The authors gratefully acknowledge the work of the many emergency department nurses who generated the triage text and of the SynSurv development team.

Conflicts of Interest

None declared.

References

1. Duffner PK, Berman PH, Fisher PG, Green J, Schneider S, Davidson C. Clinical practice guideline - neurodiagnostic evaluation of the child with a simple febrile seizure. *Am J Pediatr* 2011 Mar;127(2):389-394 [FREE Full text]
2. Sawires R, BATTERY J, Fahey M. A review of febrile seizures: recent advances in understanding of febrile seizure pathophysiology and commonly implicated viral triggers. *Front Pediatr* 2021;9:801321 [FREE Full text] [doi: [10.3389/fped.2021.801321](https://doi.org/10.3389/fped.2021.801321)] [Medline: [35096712](https://pubmed.ncbi.nlm.nih.gov/35096712/)]
3. Waruiru C, Appleton R. Febrile seizures: an update. *Arch Dis Child* 2004 Aug 01;89(8):751-756 [FREE Full text] [doi: [10.1136/adc.2003.028449](https://doi.org/10.1136/adc.2003.028449)] [Medline: [15269077](https://pubmed.ncbi.nlm.nih.gov/15269077/)]
4. Principi N, Esposito S. Vaccines and febrile seizures. *Expert Rev Vaccines* 2013 Aug 09;12(8):885-892. [doi: [10.1586/14760584.2013.814781](https://doi.org/10.1586/14760584.2013.814781)] [Medline: [23984960](https://pubmed.ncbi.nlm.nih.gov/23984960/)]
5. Chung S. Febrile seizures. *Korean J Pediatr* 2014 Oct;57(9):384-395 [FREE Full text] [doi: [10.3345/kjp.2014.57.9.384](https://doi.org/10.3345/kjp.2014.57.9.384)] [Medline: [25324864](https://pubmed.ncbi.nlm.nih.gov/25324864/)]
6. Blyth C, Currie A, Wiertsema SP, Conway N, Kirkham LA, Fuery A, et al. Trivalent influenza vaccine and febrile adverse events in Australia, 2010: clinical features and potential mechanisms. *Vaccine* 2011 Jul 18;29(32):5107-5113. [doi: [10.1016/j.vaccine.2011.05.054](https://doi.org/10.1016/j.vaccine.2011.05.054)] [Medline: [21640152](https://pubmed.ncbi.nlm.nih.gov/21640152/)]
7. Armstrong PK, Dowse GK, Effler PV, Carcione D, Blyth CC, Richmond PC, et al. Epidemiological study of severe febrile reactions in young children in Western Australia caused by a 2010 trivalent inactivated influenza vaccine. *BMJ Open* 2011 May 30;1(1):e000016 [FREE Full text] [doi: [10.1136/bmjopen-2010-000016](https://doi.org/10.1136/bmjopen-2010-000016)] [Medline: [22021725](https://pubmed.ncbi.nlm.nih.gov/22021725/)]
8. Investigation into febrile reactions in young children following 2010 seasonal trivalent influenza vaccination. Department of Health, Government of Australia. 2010. URL: <https://www.tga.gov.au/sites/default/files/alerts-medicine-seasonal-flu-100702.pdf> [accessed 2024-04-29]
9. Sweet M. Australia suspends seasonal flu vaccination of young children. *BMJ* 2010 May 04;340(may04 2):c2419. [doi: [10.1136/bmj.c2419](https://doi.org/10.1136/bmj.c2419)] [Medline: [20442237](https://pubmed.ncbi.nlm.nih.gov/20442237/)]
10. Horvath M. Review of the management of adverse events associated with Panvax and Fluvax. Department of Health, Government of Australia. 2011. URL: <https://www.tga.gov.au/news/media-releases/review-management-adverse-events-associated-panvax-and-fluvax> [accessed 2024-04-29]
11. Stokes B. Ministerial review into the public health response into the adverse events to the seasonal influenza vaccine. Minister for Health, Government of Australia. 2010. URL: https://www.health.wa.gov.au/~/-/media/Files/Corporate/Reports-and-publications/PDF/Stokes_Report.pdf [accessed 2024-04-29]
12. Clothier HJ, Crawford NW, Russell M, Kelly H, BATTERY JP. Evaluation of 'SAEFVIC', a pharmacovigilance surveillance scheme for the spontaneous reporting of adverse events following immunisation in Victoria, Australia. *Drug Saf* 2017 Jul 24;40(6):483-495. [doi: [10.1007/s40264-017-0520-7](https://doi.org/10.1007/s40264-017-0520-7)] [Medline: [28342074](https://pubmed.ncbi.nlm.nih.gov/28342074/)]
13. Clothier HJ, Lawrie J, Russell MA, Kelly H, BATTERY JP. Early signal detection of adverse events following influenza vaccination using proportional reporting ratio, Victoria, Australia. *PLoS One* 2019 Nov 1;14(11):e0224702 [FREE Full text] [doi: [10.1371/journal.pone.0224702](https://doi.org/10.1371/journal.pone.0224702)] [Medline: [31675362](https://pubmed.ncbi.nlm.nih.gov/31675362/)]
14. BATTERY JP, Clothier H. Information systems for vaccine safety surveillance. *Hum Vaccin Immunother* 2022 Dec 30;18(6):2100173 [FREE Full text] [doi: [10.1080/21645515.2022.2100173](https://doi.org/10.1080/21645515.2022.2100173)] [Medline: [36162040](https://pubmed.ncbi.nlm.nih.gov/36162040/)]

15. Triple S Project. Assessment of syndromic surveillance in Europe. *Lancet* 2011 Dec 26;378(9806):1833-1834. [doi: [10.1016/S0140-6736\(11\)60834-9](https://doi.org/10.1016/S0140-6736(11)60834-9)] [Medline: [22118433](https://pubmed.ncbi.nlm.nih.gov/22118433/)]
16. Henning KJ. What is syndromic surveillance? *MMWR Suppl* 2004 Oct 24;53:5-11 [FREE Full text] [Medline: [15714620](https://pubmed.ncbi.nlm.nih.gov/15714620/)]
17. Bawa Z, Elliot AJ, Morbey RA, Ladhani S, Cunliffe NA, O'Brien SJ, et al. Assessing the likely impact of a rotavirus vaccination program in England: the contribution of syndromic surveillance. *Clin Infect Dis* 2015 Jul 01;61(1):77-85. [doi: [10.1093/cid/civ264](https://doi.org/10.1093/cid/civ264)] [Medline: [25828997](https://pubmed.ncbi.nlm.nih.gov/25828997/)]
18. Mesfin YM, Cheng A, Lawrie J, Buttery J. Use of routinely collected electronic healthcare data for postlicensure vaccine safety signal detection: a systematic review. *BMJ Glob Health* 2019 Jul 08;4(4):e001065 [FREE Full text] [doi: [10.1136/bmjgh-2018-001065](https://doi.org/10.1136/bmjgh-2018-001065)] [Medline: [31354969](https://pubmed.ncbi.nlm.nih.gov/31354969/)]
19. Mesfin YM, Cheng AC, Enticott J, Lawrie J, Buttery JP. Use of telephone helpline data for syndromic surveillance of adverse events following immunization in Australia: a retrospective study, 2009 to 2017. *Vaccine* 2020 Jul 22;38(34):5525-5531. [doi: [10.1016/j.vaccine.2020.05.078](https://doi.org/10.1016/j.vaccine.2020.05.078)] [Medline: [32593607](https://pubmed.ncbi.nlm.nih.gov/32593607/)]
20. Hughes HE, Edeghere O, O'Brien SJ, Vivancos R, Elliot AJ. Emergency department syndromic surveillance systems: a systematic review. *BMC Public Health* 2020 Dec 09;20(1):1891 [FREE Full text] [doi: [10.1186/s12889-020-09949-y](https://doi.org/10.1186/s12889-020-09949-y)] [Medline: [33298000](https://pubmed.ncbi.nlm.nih.gov/33298000/)]
21. Rhea S, Ising A, Fleischauer AT, Deyneka L, Vaughan-Batten H, Waller A. Using near real-time morbidity data to identify heat-related illness prevention strategies in North Carolina. *J Community Health* 2012 May 1;37(2):495-500. [doi: [10.1007/s10900-011-9469-0](https://doi.org/10.1007/s10900-011-9469-0)] [Medline: [21882040](https://pubmed.ncbi.nlm.nih.gov/21882040/)]
22. Horng S, Greenbaum NR, Nathanson LA, McClay JC, Goss FR, Nielson JA. Consensus development of a modern ontology of emergency department presenting problems-the hierarchical presenting problem ontology (HaPPy). *Appl Clin Inform* 2019 May 12;10(3):409-420 [FREE Full text] [doi: [10.1055/s-0039-1691842](https://doi.org/10.1055/s-0039-1691842)] [Medline: [31189204](https://pubmed.ncbi.nlm.nih.gov/31189204/)]
23. Travers DA, Haas SW. Using nurses' natural language entries to build a concept-oriented terminology for patients' chief complaints in the emergency department. *J Biomed Inform* 2003 Aug;36(4-5):260-270 [FREE Full text] [doi: [10.1016/j.jbi.2003.09.007](https://doi.org/10.1016/j.jbi.2003.09.007)] [Medline: [14643721](https://pubmed.ncbi.nlm.nih.gov/14643721/)]
24. Picard C, Kleib M, Norris C, O'Rourke HM, Montgomery C, Douma M. The use and structure of emergency nurses' triage narrative data: scoping review. *JMIR Nurs* 2023 Jan 13;6:e41331 [FREE Full text] [doi: [10.2196/41331](https://doi.org/10.2196/41331)] [Medline: [36637881](https://pubmed.ncbi.nlm.nih.gov/36637881/)]
25. Conway M, Dowling JN, Chapman WW. Using chief complaints for syndromic surveillance: a review of chief complaint based classifiers in North America. *J Biomed Inform* 2013 Aug;46(4):734-743 [FREE Full text] [doi: [10.1016/j.jbi.2013.04.003](https://doi.org/10.1016/j.jbi.2013.04.003)] [Medline: [23602781](https://pubmed.ncbi.nlm.nih.gov/23602781/)]
26. Bouchouar E, Hetman BM, Hanley B. Development and validation of an automated emergency department-based syndromic surveillance system to enhance public health surveillance in Yukon: a lower-resourced and remote setting. *BMC Public Health* 2021 Jul 29;21(1):1247 [FREE Full text] [doi: [10.1186/s12889-021-11132-w](https://doi.org/10.1186/s12889-021-11132-w)] [Medline: [34187423](https://pubmed.ncbi.nlm.nih.gov/34187423/)]
27. Haas SW, Travers D, Waller A, Mahalingam D, Crouch J, Schwartz TA, et al. Emergency medical text classifier: new system improves processing and classification of triage notes. *Online J Public Health Inform* 2014 Oct 16;6(2):e178 [FREE Full text] [doi: [10.5210/ojphi.v6i2.5469](https://doi.org/10.5210/ojphi.v6i2.5469)] [Medline: [25379126](https://pubmed.ncbi.nlm.nih.gov/25379126/)]
28. Doan S, Maehara CK, Chaparro JD, Lu S, Liu R, Graham A, Pediatric Emergency Medicine Kawasaki Disease Research Group. Building a natural language processing tool to identify patients with high clinical suspicion for Kawasaki disease from emergency department notes. *Acad Emerg Med* 2016 May 13;23(5):628-636 [FREE Full text] [doi: [10.1111/acem.12925](https://doi.org/10.1111/acem.12925)] [Medline: [26826020](https://pubmed.ncbi.nlm.nih.gov/26826020/)]
29. Aamer H, Ofoghi B, Verspoor KM. Syndromic surveillance on the Victorian chief complaint data set using a hybrid statistical and machine learning technique. In: *Proceedings of the 2016 Health Data Analytics Conference*. 2016 Presented at: HDAC '16; October 11-12, 2016; Brisbane, Australia URL: https://ceur-ws.org/Vol-1683/hda16_aamer.pdf
30. Chapman WW, Nadkarni PM, Hirschman L, D'Avolio LW, Savova GK, Uzuner O. Overcoming barriers to NLP for clinical text: the role of shared tasks and the need for additional creative solutions. *J Am Med Inform Assoc* 2011;18(5):540-543 [FREE Full text] [doi: [10.1136/amiajnl-2011-000465](https://doi.org/10.1136/amiajnl-2011-000465)] [Medline: [21846785](https://pubmed.ncbi.nlm.nih.gov/21846785/)]
31. Lee SH, Levin D, Finley PD, Heilig CM. Chief complaint classification with recurrent neural networks. *J Biomed Inform* 2019 May;93:103158 [FREE Full text] [doi: [10.1016/j.jbi.2019.103158](https://doi.org/10.1016/j.jbi.2019.103158)] [Medline: [30926471](https://pubmed.ncbi.nlm.nih.gov/30926471/)]
32. Hsu JH, Weng TC, Wu CH, Ho TS. Natural language processing methods for detection of influenza-like illness from chief complaints. In: *Proceedings of the 2020 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*. 2020 Presented at: APSIPA-ASC '20; December 7-10, 2020; Auckland, New Zealand p. 1626-1630 URL: <https://ieeexplore.ieee.org/document/9306243> [doi: [10.23919/apsipa.2018.8659710](https://doi.org/10.23919/apsipa.2018.8659710)]
33. Goto T, Camargo Jr CA, Faridi MK, Yun BJ, Hasegawa K. Machine learning approaches for predicting disposition of asthma and COPD exacerbations in the ED. *Am J Emerg Med* 2018 Oct;36(9):1650-1654. [doi: [10.1016/j.ajem.2018.06.062](https://doi.org/10.1016/j.ajem.2018.06.062)] [Medline: [29970272](https://pubmed.ncbi.nlm.nih.gov/29970272/)]
34. Horng S, Sontag DA, Halpern Y, Jernite Y, Shapiro NI, Nathanson LA. Creating an automated trigger for sepsis clinical decision support at emergency department triage using machine learning. *PLoS One* 2017 Apr 6;12(4):e0174708 [FREE Full text] [doi: [10.1371/journal.pone.0174708](https://doi.org/10.1371/journal.pone.0174708)] [Medline: [28384212](https://pubmed.ncbi.nlm.nih.gov/28384212/)]

35. Svetcic J, Stapelberg NC, Turner K. Suicidal and self-harm presentations to emergency departments: the challenges of identification through diagnostic codes and presenting complaints. *Health Inf Manag* 2020 Jan 04;49(1):38-46. [doi: [10.1177/1833358319857188](https://doi.org/10.1177/1833358319857188)] [Medline: [31272232](https://pubmed.ncbi.nlm.nih.gov/31272232/)]
36. Ryan J. Comparison of presenting complaint vs discharge diagnosis for identifying “nonemergency” emergency department visits. *J Emerg Med* 2013 Jul;45(1):152-153. [doi: [10.1016/j.jemermed.2013.05.036](https://doi.org/10.1016/j.jemermed.2013.05.036)]
37. Singleton J, Li C, Akpunonu PD, Abner EL, Kucharska-Newton AM. Using natural language processing to identify opioid use disorder in electronic health record data. *Int J Med Inform* 2023 Mar;170:104963 [FREE Full text] [doi: [10.1016/j.ijmedinf.2022.104963](https://doi.org/10.1016/j.ijmedinf.2022.104963)] [Medline: [36521420](https://pubmed.ncbi.nlm.nih.gov/36521420/)]
38. Moore K, Black J, Rowe S, Franklin L. Syndromic surveillance for influenza in two hospital emergency departments. Relationships between ICD-10 codes and notified cases, before and during a pandemic. *BMC Public Health* 2011 May 18;11(1):338 [FREE Full text] [doi: [10.1186/1471-2458-11-338](https://doi.org/10.1186/1471-2458-11-338)] [Medline: [21592398](https://pubmed.ncbi.nlm.nih.gov/21592398/)]
39. Rozova V, Witt K, Robinson J, Li Y, Verspoor K. Detection of self-harm and suicidal ideation in emergency department triage notes. *J Am Med Inform Assoc* 2022 Jan 29;29(3):472-480 [FREE Full text] [doi: [10.1093/jamia/ocab261](https://doi.org/10.1093/jamia/ocab261)] [Medline: [34897466](https://pubmed.ncbi.nlm.nih.gov/34897466/)]
40. Spasic I, Nenadic G. Clinical text data in machine learning: systematic review. *JMIR Med Inform* 2020 Mar 31;8(3):e17984 [FREE Full text] [doi: [10.2196/17984](https://doi.org/10.2196/17984)] [Medline: [32229465](https://pubmed.ncbi.nlm.nih.gov/32229465/)]
41. Valman HB. ABC of one to seven. Febrile convulsions. *BMJ* 1993 Jul 26;306(6894):1743-1745 [FREE Full text] [doi: [10.1136/bmj.306.6894.1743](https://doi.org/10.1136/bmj.306.6894.1743)] [Medline: [8257494](https://pubmed.ncbi.nlm.nih.gov/8257494/)]
42. Chang TA, Bergen BK. Language model behavior: a comprehensive survey. arXiv Preprint posted online March 20, 2023 [FREE Full text] [doi: [10.1162/coli_a_00492](https://doi.org/10.1162/coli_a_00492)]
43. Khademi S, Palmer C, Dimaguila GL, Javed M, Buttery J, Black J. Data augmentation to improve syndromic detection from emergency department notes. In: Proceedings of the 2023 Australasian Computer Science Week. 2023 Presented at: ACSW '23; January 30-February 3, 2023; Melbourne, Australia p. 198-205 URL: <https://dl.acm.org/doi/abs/10.1145/3579375.3579401> [doi: [10.1145/3579375.3579401](https://doi.org/10.1145/3579375.3579401)]
44. Lewis P, Ott M, Du J, Stoyanov V. Pretrained language models for biomedical and clinical tasks: understanding and extending the state-of-the-art. In: Proceedings of the 3rd Clinical Natural Language Processing Workshop. 2020 Presented at: ClinicalNLP '20; November 19, 2020; Virtual Event p. 146-157 URL: <https://aclanthology.org/2020.clinicalnlp-1.17.pdf> [doi: [10.18653/v1/2020.clinicalnlp-1.17](https://doi.org/10.18653/v1/2020.clinicalnlp-1.17)]
45. Khademi Habibabadi S, Delir Haghighi P, Burstein F, Buttery J. Vaccine adverse event mining of twitter conversations: 2-phase classification study. *JMIR Med Inform* 2022 Jul 16;10(6):e34305 [FREE Full text] [doi: [10.2196/34305](https://doi.org/10.2196/34305)] [Medline: [35708760](https://pubmed.ncbi.nlm.nih.gov/35708760/)]
46. Velupillai S, Suominen H, Liakata M, Roberts A, Shah AD, Morley K, et al. Using clinical natural language processing for health outcomes research: overview and actionable suggestions for future advances. *J Biomed Inform* 2018 Dec;88:11-19 [FREE Full text] [doi: [10.1016/j.jbi.2018.10.005](https://doi.org/10.1016/j.jbi.2018.10.005)] [Medline: [30368002](https://pubmed.ncbi.nlm.nih.gov/30368002/)]
47. Lederman A, Lederman R, Verspoor K. Tasks as needs: reframing the paradigm of clinical natural language processing research for real-world decision support. *J Am Med Inform Assoc* 2022 Oct 12;29(10):1810-1817 [FREE Full text] [doi: [10.1093/jamia/ocac121](https://doi.org/10.1093/jamia/ocac121)] [Medline: [35848784](https://pubmed.ncbi.nlm.nih.gov/35848784/)]
48. Zhang Z, Strubell E, Hovy E. A survey of active learning for natural language processing. In: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. 2022 Presented at: EMNLP '22; December 7-11, 2022; Abu Dhabi, United Arab Emirates p. 6166-6190 URL: <https://aclanthology.org/2022.emnlp-main.414.pdf> [doi: [10.18653/v1/2022.emnlp-main.414](https://doi.org/10.18653/v1/2022.emnlp-main.414)]

Abbreviations

- AEFI:** adverse event following immunization
- BERT:** Bidirectional Encoder Representations from Transformers
- BiGRU:** bidirectional gated recurrent unit
- CNN:** convolutional neural network
- CPU:** central processing unit
- ED:** emergency department
- GPU:** graphics processing unit
- GRU:** gated recurrent unit
- IDF:** inverse document frequency
- LSTM:** long short-term memory
- ML:** machine learning
- NLP:** natural language processing
- ONNX:** Open Neural Network Exchange
- SGD:** stochastic gradient descent
- SQL:** structured query language

SVC: support vector classification

XG Boost: extreme gradient boosting

Edited by K El Emam, B Malin; submitted 09.11.23; peer-reviewed by LH Yao, S Ma; comments to author 13.01.24; revised version received 09.03.24; accepted 30.03.24; published 30.08.24.

Please cite as:

Khademi S, Palmer C, Javed M, Dimaguila GL, Clothier H, Buttery J, Black J

Near Real-Time Syndromic Surveillance of Emergency Department Triage Texts Using Natural Language Processing: Case Study in Febrile Convulsion Detection

JMIR AI 2024;3:e54449

URL: <https://ai.jmir.org/2024/1/e54449>

doi: [10.2196/54449](https://doi.org/10.2196/54449)

PMID:

©Sedigh Khademi, Christopher Palmer, Muhammad Javed, Gerardo Luis Dimaguila, Hazel Clothier, Jim Buttery, Jim Black. Originally published in JMIR AI (<https://ai.jmir.org>), 30.08.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR AI, is properly cited. The complete bibliographic information, a link to the original publication on <https://www.ai.jmir.org/>, as well as this copyright and license information must be included.

Original Paper

Enhancing Clinical Relevance of Pretrained Language Models Through Integration of External Knowledge: Case Study on Cardiovascular Diagnosis From Electronic Health Records

Qiu hao Lu^{1,2,3}, PhD; Andrew Wen^{1,2}, MS; Thien Nguyen³, PhD; Hongfang Liu^{1,2}, PhD

¹McWilliams School of Biomedical Informatics, University of Texas Health Science Center, Houston, TX, United States

²Department of AI and Informatics, Mayo Clinic, Rochester, MN, United States

³Department of Computer Science, University of Oregon, Eugene, OR, United States

Corresponding Author:

Hongfang Liu, PhD

McWilliams School of Biomedical Informatics

University of Texas Health Science Center

7000 Fannin Street

Houston, TX, 77030

United States

Phone: 1 713 500 4472

Email: Hongfang.Liu@uth.tmc.edu

Abstract

Background: Despite their growing use in health care, pretrained language models (PLMs) often lack clinical relevance due to insufficient domain expertise and poor interpretability. A key strategy to overcome these challenges is integrating external knowledge into PLMs, enhancing their adaptability and clinical usefulness. Current biomedical knowledge graphs like UMLS (Unified Medical Language System), SNOMED CT (Systematized Medical Nomenclature for Medicine–Clinical Terminology), and HPO (Human Phenotype Ontology), while comprehensive, fail to effectively connect general biomedical knowledge with physician insights. There is an equally important need for a model that integrates diverse knowledge in a way that is both unified and compartmentalized. This approach not only addresses the heterogeneous nature of domain knowledge but also recognizes the unique data and knowledge repositories of individual health care institutions, necessitating careful and respectful management of proprietary information.

Objective: This study aimed to enhance the clinical relevance and interpretability of PLMs by integrating external knowledge in a manner that respects the diversity and proprietary nature of health care data. We hypothesize that domain knowledge, when captured and distributed as stand-alone modules, can be effectively reintegrated into PLMs to significantly improve their adaptability and utility in clinical settings.

Methods: We demonstrate that through adapters, small and lightweight neural networks that enable the integration of extra information without full model fine-tuning, we can inject diverse sources of external domain knowledge into language models and improve the overall performance with an increased level of interpretability. As a practical application of this methodology, we introduce a novel task, structured as a case study, that endeavors to capture physician knowledge in assigning cardiovascular diagnoses from clinical narratives, where we extract diagnosis-comment pairs from electronic health records (EHRs) and cast the problem as text classification.

Results: The study demonstrates that integrating domain knowledge into PLMs significantly improves their performance. While improvements with ClinicalBERT are more modest, likely due to its pretraining on clinical texts, BERT (bidirectional encoder representations from transformer) equipped with knowledge adapters surprisingly matches or exceeds ClinicalBERT in several metrics. This underscores the effectiveness of knowledge adapters and highlights their potential in settings with strict data privacy constraints. This approach also increases the level of interpretability of these models in a clinical context, which enhances our ability to precisely identify and apply the most relevant domain knowledge for specific tasks, thereby optimizing the model's performance and tailoring it to meet specific clinical needs.

Conclusions: This research provides a basis for creating health knowledge graphs infused with physician knowledge, marking a significant step forward for PLMs in health care. Notably, the model balances integrating knowledge both comprehensively and selectively, addressing the heterogeneous nature of medical knowledge and the privacy needs of health care institutions.

KEYWORDS

knowledge integration; pre-trained language models; physician reasoning; adapters; physician; physicians; electronic health record; electronic health records; EHR; healthcare; heterogeneous; healthcare institution; healthcare institutions; proprietary information; healthcare data; methodology; text classification; data privacy; medical knowledge

Introduction

Background

In recent years, pretrained language models (PLMs) have revolutionized many areas of natural language processing (NLP), demonstrating proficiency in handling a broad spectrum of general-domain text tasks. However, their performance declines when confronted with specialized domains, such as health care, as clinical text often presents unique linguistic characteristics and semantics that differ from standard language [1,2]. The extensive proliferation of electronic health records (EHRs) further underscores the gap, highlighting the demand for domain-specific methods in PLMs.

Although there are domain-specific PLMs designed by training on large-scale clinical data sets, they often fail to capture the depth and breadth of knowledge scattered across diverse biomedical sources [3]. This limitation calls for an approach that integrates specific domain knowledge into PLMs, enhancing their effectiveness and accuracy in specialized contexts.

To address this need, we propose a dual strategy—the strategic incorporation of external knowledge from diverse sources in a unified yet compartmentalized manner. Biomedical domain knowledge is inherently heterogeneous and stored in a variety of formats. A unified approach that simultaneously incorporates various knowledge sources is essential to manage this diversity. Traditional methods of sequential training with new knowledge sources are inefficient and risk losing previously integrated knowledge due to continuous model parameter adjustments. A unified model overcomes these challenges by integrating diverse knowledge without the need for repeated, individualized retraining.

Furthermore, the broad diversity of domain knowledge sources, each relevant to different tasks in its own way, underscores the need for a compartmentalized approach. This strategy allows for the selective integration of the most relevant knowledge, avoiding information overload. In addition, given that each institution manages its own proprietary repository of data and knowledge, often governed by protected health information (PHI) regulations, a method that potentially respects institutional boundaries is desirable. This could enable an institution to freely choose to equip a widely shareable foundational model with its particular data, thereby enabling an adaptable and compliant framework that can cater to diverse institutional needs without compromising data privacy and security.

Building on this rationale, we introduce a specific case study in the cardiovascular domain to demonstrate our approach. This involves extracting diagnosis-comment pairs from EHRs and approaching the problem through text classification, predicting diagnoses based on physician comments. Essentially, we take PLMs with a linear head on top as the foundational prediction

model and fine-tune them on this specific task, optimizing it to better capture the specialized knowledge and clinical terminologies present in physician comments within the cardiovascular domain.

As most clinical PLMs, such as clinical bidirectional encoder representations from transformer (ClinicalBERT) [4], are primarily trained on large-scale free texts and lack integration with structured domain knowledge, they often demonstrate suboptimal performance in knowledge-driven tasks [5-7]. To address this limitation, we incorporate the Diverse Adapters for Knowledge Integration (DAKI) framework [6] for knowledge infusion, which integrates domain knowledge adaptively from multiple sources. More specifically, we train 3 distinct adapters, each tailored to encapsulate domain knowledge from a specific source, that are (1) the Unified Medical Language System (UMLS) Metathesaurus, (2) Wikipedia articles, and (3) semantic grouping information for biomedical concepts. This approach effectively augments PLMs, enhancing their performance within the clinical context. The adapter-enhanced PLMs retain a unified utility, functioning as standard PLMs, while simultaneously featuring a compartmentalized structure, where adapters are incorporated in a plug-and-play manner, ensuring flexibility and transferability. The contributions can be summarized as follows: (1) we propose a novel task aimed at capturing physician knowledge in the cardiovascular domain through text classification of diagnosis-comment pairs from EHRs. The encouraging performance of our models on this task validates its feasibility, demonstrating the potential of PLMs in capturing medical insights. (2) Upon integrating domain knowledge through the DAKI framework, the models not only exhibit enhanced performance but also an increased level of interpretability, where we can closely examine and clarify which external domain knowledge is activated during tasks. Such interpretability could further enable the identification of vital knowledge pieces, refine the fine-tuning of models for particular tasks, and assist in adjusting the applied domain knowledge to be more task-specific. (3) The domain knowledge demonstrates transferability when injecting respective adapters into different PLMs, where pretrained knowledge adapters also prove effective when equipped with other, previously unseen PLMs. This highlights the potential for heterogeneous knowledge infusion while considering institutional boundaries, laying a foundational step toward the development of health knowledge graphs enriched with physician knowledge.

Related Work

Patient diagnosis prediction is a challenging task due to the complex and knowledge-intensive nature of this field. Most existing studies heavily rely on codified, numerical, or time-series features of patients, where significant features are manually selected as input to downstream machine learning models. Franz et al [8] extracted all numerical observations

from MIMIC-III (Medical Information Mart for Intensive Care III) data set [9], for example, vital sign measures and lab results, and fed them as input into a 4-layer neural network (1 convolutional neural network [10] layer spanning across the time dimension followed by 3 fully connected layers) for multiclass classification. Zoabi et al [11] selected a set of features including sex, age, symptoms (cough, fever, sore throat, shortness of breath, and headache), and known contact as input and fed them into a gradient-boosting machine model to track COVID-19. Meanwhile, with the rapid growth of NLP techniques, researchers have been exploring the clinical notes of EHRs for a wide variety of clinically relevant tasks, including diagnosis prediction. For example, Franz et al [8] fine-tuned ClinicalBERT [12] for the prediction and significantly outperformed their numerical method. Another line of research aimed to leverage the multimodality of EHRs as there exists rich structural information within EHRs, for example, the interactions among users, symptoms, and diseases [13], where these interactions are captured through encoding EHRs through graph neural networks [14,15]. The task of this work differs from the aforementioned studies in that we only use a single piece of physician comment as input, and instead of pushing state-of-the-art predictive performance, we try to understand the insight of a physician by capturing their reasoning on the diagnosis.

While PLMs excel on general-domain text, their performance over domain-specific text is relatively poor due to domain shift [2]. In the last few years, several domain-specific PLMs have been proposed to mitigate the issue, for example, BioBERT [16], ClinicalBERT [12], ClinicalBERT [4], PubMedBERT [17], ClinicalT5 [18], etc. Despite their specificity, training these models demands significant time and resources. Moreover, recent findings indicate that even these specialized models can struggle in certain scenarios, particularly when reliable knowledge retrieval is essential for complex domain-specific reasoning [3].

Beyond acquiring domain knowledge through pretraining, a distinct research trajectory emphasizes knowledge infusion, wherein domain knowledge is intentionally injected into language models [6,7,19-23]. Typically, this involves adding an auxiliary training objective driven by knowledge. This approach facilitates additional pretraining or fine-tuning of existing models, thereby cutting down on training expenses, though it can still demand significant resources. For instance, Wang et al [7] jointly optimized language modeling with a knowledge embedding objective. Zhang et al [23] fused PLMs with graph neural networks through layered modality interactions, enabling bidirectional information flow for enhanced reasoning in question-answering tasks. Our choice to use DAKI [6] for knowledge infusion is motivated by 3 principal reasons, that are (1) the framework integrates domain knowledge of varied sources and formats, which reflects the heterogeneous nature of the domain knowledge; (2) focusing on training adapters, instead of the entire language model, presents a more sustainable and efficient approach; and (3) the knowledge adapters are integrated in a plug-and-play manner that increases both flexibility and interpretability.

Proposed Task Design

Data Collection and Structure

The experiment was conducted using clinical notes generated by the Mayo Clinic Rochester Campus between January 1 and December 31, 2015, corresponding to roughly 5 million documents. Specifically, we extracted the problem entries from the Impression/Report/Plan (IRP) section in the clinical note as it contained a diagnostic problem list that was used to summarize the main findings [24]. The entries are recorded as numbered items and each item is a textual description of the diagnosis followed by a physician comment detailing their reasoning for giving a diagnosis. We then convert them into <entity, comment> pairs by mapping the textual descriptions of diagnosis to entities and associated UMLS concept unique identifiers (CUIs) using SciSpacy [25]. We specifically perform entity linking for diseases and syndromes, in light of the observation that medical interests arise primarily around symptoms and problems [26]. After filtering to only clinical narratives generated in the Department of Cardiovascular Medicine and removing unrecognized or unlinkable texts, 174,980 valid pairs were generated corresponding to 30,240 patients. We then split the data into 10 folds where 8 folds for training, 1 fold for development, and 1 fold for testing were at the patient level.

Task Objective and Metrics

The task is cast as a multiclass text classification problem, that is, to predict the assigned diagnosis (entity) from a physician's comment detailing their reasoning for assigning a diagnosis. As most (linked) entities occur only once in the prepared data set, we use the most frequent top 50 entities as the targets for all experiments in this study. For instance, the top 10 most frequent entities that appear in the training set are "hypertensive disease," "hyperlipidemia," "sleep apnea, obstructive," "atrial fibrillation," "coronary arteriosclerosis," "hypothyroidism," "diabetes mellitus, non-insulin-dependent," "gastroesophageal reflux disease," "chronic kidney diseases," and "dyslipidemias." We use the top-k (k=1,3,5,10) accuracy classification score as the evaluation metric, which computes the number of times where the correct label is among the top-k labels predicted (ranked by predicted scores).

Methods

We consider 2 prediction models in this study, that are, the PLMs and those equipped with DAKI [6].

Foundational Models

For the foundational prediction models, we used BERT-base-uncased [27], ALBERT-xxlarge-v2 [28], and ClinicalBERT-base [4] to cover base or large, and general or specific domain variants. Essentially, we encode the physician comment with the models and feed the average pooled representations into a linear layer for prediction. The model is fine-tuned by optimizing a cross-entropy loss.

Models Equipped With DAKI

To facilitate prediction on clinical text, we leverage a novel framework that incorporates DAKI into PLMs. The adapter in this framework is a small bottleneck feed-forward network with

a residual connection that is placed within PLMs, as illustrated in Figure 1. One can also incorporate a more advanced adapter structure, such as the LoRA (low-rank) adapter [29]. This framework consists of 3 major components, which are the base PLM, pretrained knowledge-specific adapters, and the knowledge controller (CTRL) that adaptively activates the adapters, as illustrated in Figure 2. Generally, when pretraining a knowledge adapter, the parameters of the base PLM are frozen, and only the adapter is optimized. In this way, we inject specific knowledge into an adapter. By equipping PLMs with adapters, one can inject domain knowledge into the models without touching the original parameters of PLMs, enhancing their representation capabilities on domain-specific text. Essentially, a knowledge adapter is independently pretrained to encode domain knowledge, and the trained adapters are then plugged into DAKI for downstream fine-tuning, where the knowledge adapters are adaptively activated by the knowledge controller. Therefore, the usage of DAKI is simple and straightforward as the output can be considered as the last hidden states of a PLM.

We use the best version of ALBERT (ie, ALBERT-xxlarge-v2 [28]) as the base PLM for adapter pretraining. In this study, we incorporate 3 clinically relevant knowledge adapters that integrate disparate domain knowledge from the UMLS Metathesaurus (knowledge graph adapter [KG]), the Wikipedia articles for diseases (disease adapter [DS]), and the semantic

groupings (semantic grouping adapter [SG]). More specifically, the KG captures relational patterns within medical entities using the UMLS Metathesaurus. It is trained on triples from UMLS, treated as textual sequences, to predict the plausibility of these relational statements. For the DS, disease-related textual descriptions are sourced from Wikipedia, with the training process focusing on inferring disease names through masked language modeling, enhancing the model's grasp on disease contexts. The SG uses UMLS semantic groupings to predict the categorization of medical concepts, leveraging textual definitions to understand and classify medical concepts into coherent groupings. Essentially, we try to enforce the KG to capture the relationships between medical entities, the DS to help PLMs understand the definitions and contexts for diseases, and the SG to maintain semantic coherence within a categorization group. We refer the readers to our previous work for a more detailed treatment of the architecture and training objectives of DAKI [6].

We equip the 3 foundational models, that is, BERT, ALBERT, and ClinicalBERT, with DAKI, respectively, and enable all the previously trained adapters within the framework for the experiments. Likewise, we encode the physician comments with the DAKI models and feed the average pooled representations to a linear layer for prediction.

Figure 1. Adapter module, for example, a bottleneck feed-forward network with a residual connection.

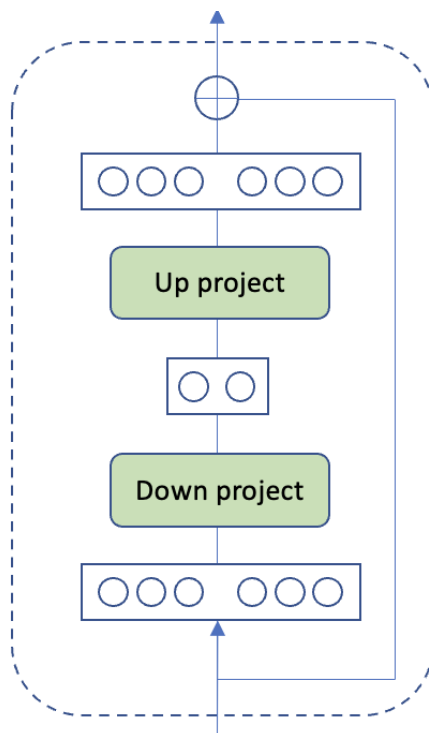
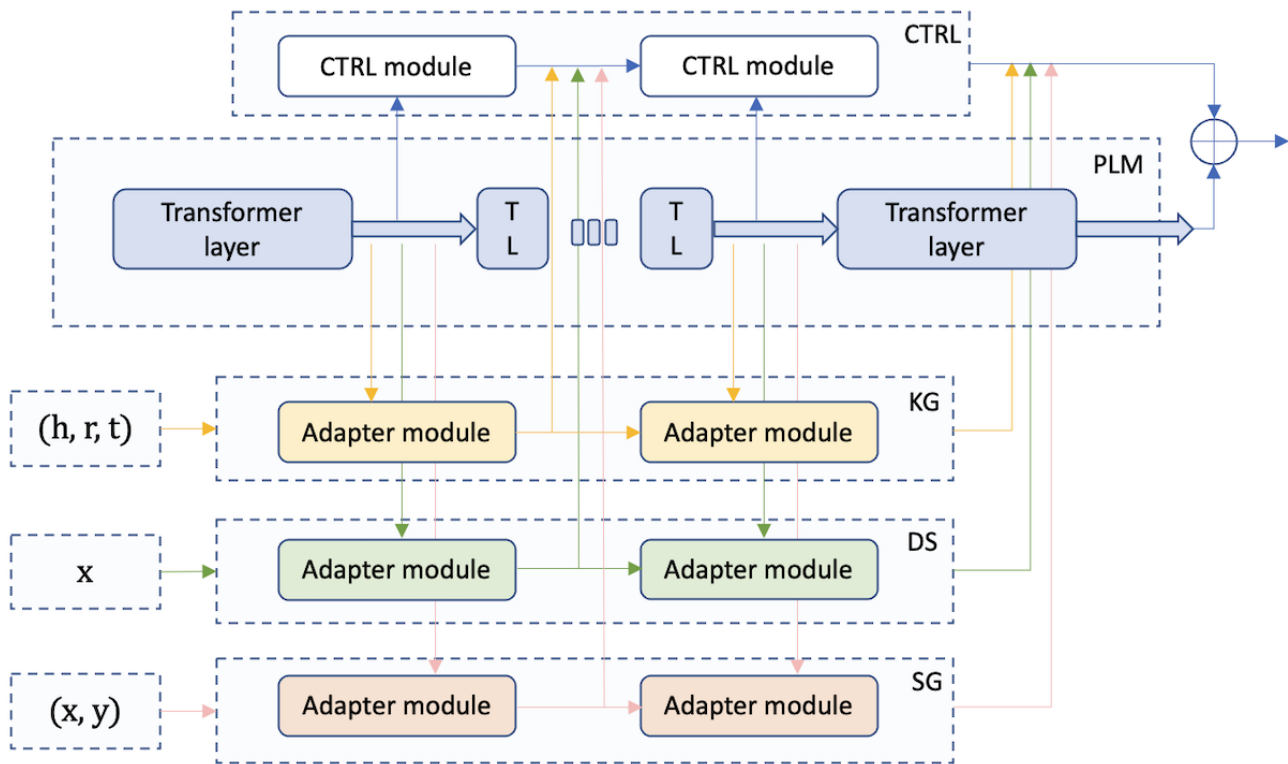


Figure 2. Architecture of DAKI [6]. CTRL: knowledge controller; DAKI: Diverse Adapters for Knowledge Integration; DS: disease adapter; h, r, t: head, relation, tail; KG: knowledge graph adapter; PLM: pretrained language model; SG: semantic grouping adapter; TL: transformer layer.



Ethical Considerations

We used the Mayo Clinic IRP data, and this study was approved by the Mayo Clinic Institutional Review Board (#20-001137) for human participants research. The data were not anonymous. No compensation was offered to participants in the study. Due to the presence of private health information in the clinical data set, we do not distribute any recordings or models trained on these recordings. Access to the clinical data is restricted to Mayo Clinic researchers who have the appropriate authorization.

Results

Overview

We present the performance of models both with and without DAKI on the test set and development set in Table 1. Generally, all the models are fine-tuned on the development set, and the best epochs of that are selected to report their performance on the test set. The results indicate that the infusion of domain knowledge into PLMs through DAKI consistently boosts their overall performance. Notably, DAKI-ALBERT demonstrates compelling performance gain over ALBERT across all the metrics, compared with the other 2 foundational models, which

is almost expected as the adapters are trained with ALBERT as the base PLM. On the other hand, the improvement with ClinicalBERT is comparatively slight, and we hypothesize that this is due to ClinicalBERT's extensive exposure to clinical text during its pretraining, rendering DAKI-ClinicalBERT less striking.

Another key observation from our results is that DAKI-BERT not only matches but in certain metrics surpasses the performance of ClinicalBERT. This highlights the advantages of incorporating knowledge adapters, particularly given that DAKI-BERT achieves such results without needing extensive and sensitive clinical text corpora. Such transferability of knowledge adapters also indicates a potential for heterogeneous knowledge infusion while respecting institutional boundaries, especially in contexts where each institution possesses its own exclusive data repository due to PHI constraints.

Moreover, considering it is a complex 50-class classification problem, these results are commendably robust. They not only shed light on the feasibility of encapsulating physician reasoning but also highlight the potential of transferable or portable domain knowledge.

Table 1. Overall performance.

Data sets and metrics	Test				Development			
	Acc@1 ^a	Acc@3	Acc@5	Acc@10	Acc@1	Acc@3	Acc@5	Acc@10
Without DAKI^b, %								
BERT	48.81	69.7	76.73	85.37	50.56	70.21	76.84	84.92
ALBERT	48.02	69.35	76.7	85.82	50.47	69.37	76	84.6
ClinicalBERT	48.49	69.88	77.05	85.96	50.63	70.27	76.87	84.97
With DAKI, %								
BERT	48.2	70.26	77.64	86.38	50.59	70.19	76.77	84.78
ALBERT	48.32	69.93	77.6	86.3	50.7	70.48	76.82	84.84
ClinicalBERT	48.73	70.15	77.05	85.94	51.14	69.99	76.45	84.75

^aAcc@k: the number of times where the correct label is among the top-k labels predicted (ranked by predicted scores).

^bDAKI: Diverse Adapters for Knowledge Integration.

Ablation Study

To investigate the influence of each of the knowledge adapters, we conduct an ablation study and show the results in Table 2. We take DAKI-BERT and DAKI-ClinicalBERT for comparison as they have the same number of parameters. We gradually remove the knowledge adapters from the complete setting (ie, all 3 equipped) and this makes 6 conditions, as shown in the table. Essentially for DAKI-BERT, the results of the ablated models demonstrate varying degrees of decline in performance,

indicating the necessity of each source of external knowledge. For DAKI-ClinicalBERT, however, the situation is different. When 1 knowledge adapter is removed (ie, KG or DS), the performance gets improved, which is consistent with our conjecture that ClinicalBERT has been exposed to clinical knowledge during pretraining and this weakens the knowledge adapters' impact. When 2 knowledge adapters are removed, the performance gets decreased at a lower level compared with that of DAKI-BERT, indicating the effectiveness and complementary nature of the knowledge adapters.

Table 2. Ablation analysis on the test set.

Ablated model and metrics	DAKI ^a -BERT					DAKI-ClinicalBERT				
	Acc@1 ^b	Acc@3	Acc@5	Acc@10	Δ^c	Acc@1	Acc@3	Acc@5	Acc@10	Δ
Baseline (no ablation), %										
With all	48.2	70.26	77.64	86.38	— ^d	48.73	70.15	77.05	85.94	—
1 adapter removed, %										
KG ^e	48.04	69.57	77.15	85.85	-1.87	48.65	70.16	77.77	86.8	1.50
DS ^f	47.62	70.51	77.32	85.85	-1.18	48.49	69.87	77.78	86.41	0.67
SG ^g	48.49	69.91	77.51	86.39	-0.17	48.65	69.49	76.91	86.22	-0.60
2 adapters removed, %										
KG-DS	48.39	69.66	76.77	85.97	-1.68	48.25	69.28	77.32	86.23	-0.80
KG-SG	48.83	69.93	76.63	85.81	-1.68	48.52	69.59	77.22	86.18	-0.36
DS-SG	47.55	69.63	76.94	86.06	-2.30	48.03	69.67	77.47	86.39	-0.31

^aDAKI: Diverse Adapters for Knowledge Integration.

^bAcc@k: the number of times where the correct label is among the top-k labels predicted (ranked by predicted scores).

^c Δ : the change of accumulated accuracy scores.

^dNot applicable.

^eKG: knowledge graph adapter.

^fDS: disease adapter.

^gSG: semantic grouping adapter.

Analysis

In this section, we want to analyze and answer the following research questions: (1) what knowledge is lacking in the PLMs for the physician reasoning task? (2) How do the models perform on different target diagnoses, that is, at the individual level? (3) How does the knowledge affect the representations at the token level?

Knowledge Activation

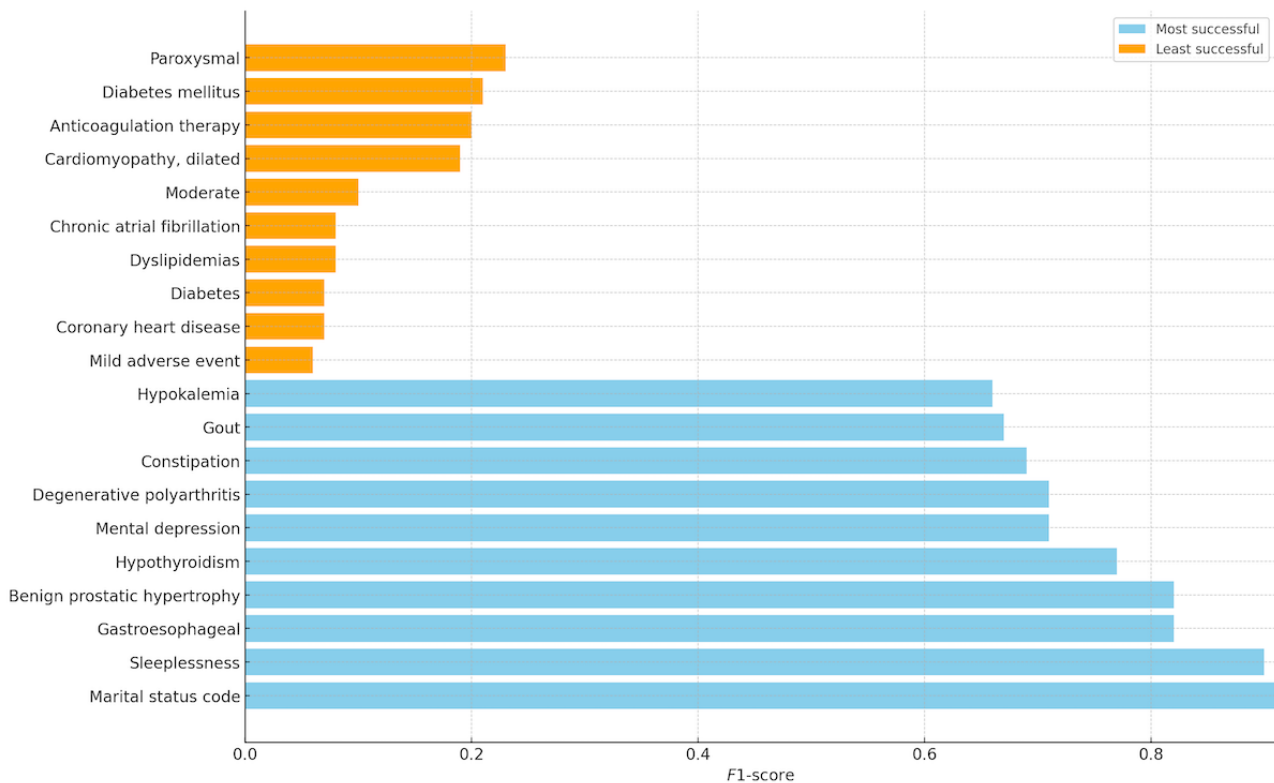
Due to DAKI's inherent flexibility, we are able to provide a high-level representation of the adapter activations during the inference process. As depicted in [Multimedia Appendix 1](#), we compute the softmax activations across 3 key layers within the encoders of DAKI-BERT (left) and DAKI-ClinicalBERT (right) where the adapters are situated. These activations are then averaged across all test set instances. Notably, the disease knowledge consistently stands out in its importance and activation across these layers, when compared with the other 2 knowledge types. We conjecture that the specificity and

relevance of DS to the model's tasks allow it to have a more significant influence on the encoder's activation patterns. This reinforces the notion that domain-specific knowledge, particularly when closely aligned with the predictive tasks, is crucial for the model's decision-making process. The injected knowledge also demonstrates a more pronounced effect on BERT than on ClinicalBERT. This distinction is likely because ClinicalBERT has previously encountered clinical data sets during its pretraining phase, which aligns with the observations detailed in [Table 1](#). The diminished reliance of ClinicalBERT on the knowledge adapters underscores the importance of identifying knowledge that truly complements specific PLMs.

Impact Pattern

We also investigate the impact pattern of these knowledge adapters. Essentially, we show the top 10 most and least successful targets in [Figure 3](#). We also observe that the targets with the biggest improvement are among the least successful targets, as shown in [Multimedia Appendix 2](#). The impact is evaluated in terms of the F_1 -score.

Figure 3. Understanding performance variability: most versus least successful targets.



In general, we believe targets that demand more tests to diagnose are easier to predict, for example, Gout. Such targets might demonstrate more unique textual context in the comment that facilitates the prediction. On the other hand, targets that are easier to diagnose are more challenging for the model to identify. For instance, it makes more sense to diagnose “Diabetes” with “150 mg/dL” than with “blood sugar.” Moreover, we observe that nearly half (ie, 4 out of 10) of the most-impacted targets are among the least successful ones (ie, represented in orange in [Multimedia Appendix 2](#)). This pattern underscores the utility of the domain knowledge we have incorporated into the PLMs. It indicates that this specialized knowledge is particularly

effective for enhancing the model's capability to accurately predict outcomes for what are considered harder targets. This suggests that targeted interventions in the training process can yield substantial improvements in predictive accuracy.

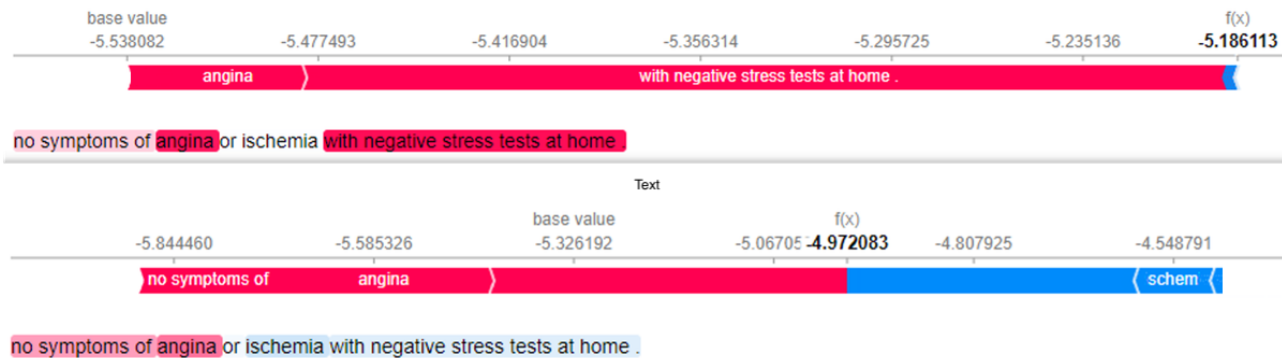
Contribution Shift

In the end, we would like to understand how the injected knowledge affects PLMs in the specific task. We use the SHapley Additive exPlanations (SHAP) tool, a game theoretic approach to explaining the output of any machine learning model [30], to explain the results. We take one of the hardest targets, that is, coronary heart disease, as an example and investigate

the contribution distribution of tokens from ClinicalBERT and DAKI-ClinicalBERT, as shown in Figure 4. Red means positive contribution (ie, predicting the comment to be coronary heart disease among all the targets in this case), and blue means negative contribution. The $f(x)$ is the model's score for this observation, where a higher score leads the model to predict the

specific class. Essentially, we observe that with the external knowledge, the DAKI-ClinicalBERT model is more sensitive to the tokens and their contribution to the prediction, compared with ClinicalBERT that treats the tokens almost equally. Such contribution shift indicates that the injected knowledge helps PLMs capture the semantics of text at the token level.

Figure 4. Contribution shift analysis of ClinicalBERT (top) and DAKI-ClinicalBERT (bottom). Darker shades of pink indicate a positive contribution and the shades of blue indicate a negative contribution to the target, that is, coronary heart disease. DAKI: Diverse Adapters for Knowledge Integration.



Discussion

Principal Findings

Interpretability is a major issue in machine learning, especially in the clinical setting. The reasons are 2-fold. First, it is essential for physicians to understand how a model is making its predictions in order to trust and effectively use the model. This is particularly important in the medical field because the consequences of incorrect predictions can be severe. Second, machine learning techniques, especially deep learning models, are hard to interpret, which makes it difficult for physicians to identify potential biases or errors in the model. To improve the interpretability in the application of machine learning to the clinical setting, we consider constructing a health knowledge graph so that the models are used responsibly and that the consequences of incorrect predictions are minimized.

In recent years, there has been a surge of interest in creating and using external health knowledge graphs to enhance the domain adaptation and interpretability of PLMs. An optimal health knowledge graph can be used for a variety of purposes, such as, (1) for research purposes, they could be used to represent the relationships between different medical conditions, treatments, and patient characteristics that a physician considers when deciding on a course of treatment for a patient and this could help to clarify the reasoning behind the decision and identify any factors that may have influenced the decision; (2) for analysis purposes, they could help to identify patterns and factors in physicians' decision-making process, which can be important for improving the quality and efficiency of hospital care; and (3) for practical purposes, such graphs could support clinical decision-making by providing physicians with information and guidance to help them make informed decisions about patient care.

Nevertheless, traditional biomedical knowledge graphs, including the UMLS [31], the Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT) [5], the Human Phenotype Ontology (HPO) [32], etc, mostly consist of

biomedical concepts and their relationships, along with textual descriptions, and can struggle to fulfill the third purpose, that is, clinical practice. The reason lies in their limited capacity to incorporate practical physician knowledge, which is crucial for clinical applications.

Physician knowledge is a key item of interest for inclusion in health knowledge graphs, as mining a health knowledge graph (as opposed to manual construction) provides the potential for discovering latent clinical knowledge that may not be self-evident. Such items can be found within physician reasoning behind assigning a diagnosis, as such diagnoses are typically made based on an application of the individual physician's knowledge. As physician reasoning is primarily not encoded in structured data forms, we must instead turn to NLP techniques, for example, the PLMs, on clinical narratives, which can include symptom descriptions, reasons for diagnosis, patient activities, and patient histories with the aim of helping physicians express a holistic picture of the patient [8].

As a preliminary step, we aim to explore and model the thought process and decision-making in the clinic by capturing physician reasoning. With the experiment of text-based diagnosis prediction, we believe that foundational PLMs are capable of capturing physician knowledge given relatively high performance in top-k ($k=1,3,5,10$) returned results, especially when compared with random chance.

Moreover, by injecting external domain knowledge from 3 disparate sources (ie, the UMLS Metathesaurus, the Wikipedia articles, and the semantic groupings) into the PLMs through adapters, we show that the models' performance gets consistently improved with an increased level of interpretability. Essentially, the framework's flexibility enables us to investigate and interpret what external domain knowledge is activated and how it contributes to the model in capturing physician reasoning.

The transferability of the knowledge adapters is also notably highlighted. As demonstrated in Table 1, DAKI-BERT's performance, on par with ClinicalBERT, underscores the

adapters' transferability, given that DAKI-BERT achieves comparable performance even without using extensive and confidential clinical text corpora. This transferability implies the potential for infusing heterogeneous knowledge while honoring institutional boundaries; institutions can create their adapters using proprietary data and knowledge, thereby sharing only the foundational models and not the institution-specific adapters. This aspect warrants further exploration in future studies.

As for future work, we would further investigate the impact of different sources of knowledge over individual diagnosis, that is, how external knowledge affects the judgments over the 50 diagnoses. We would also explore and incorporate other sources of knowledge such as the ontological structure of the target diagnoses. By combining the internal knowledge within the EHRs and the external knowledge accumulated throughout knowledge bases under a unified framework, we would improve the interpretability of machine learning models in the clinical scenario and facilitate the construction of a health knowledge graph eventually.

Although the experiments demonstrate the effectiveness of our method, there are still some limitations that can be improved. First, the impact of knowledge adapters over different clinically relevant tasks remains unclear as only one task is considered in

this work. Second, the range of external knowledge is a bit limited, for example, the inherent ontological structure of the targets is not leveraged, as mentioned above. Third, there is a lack of clinical explanation for the observations at an individual level, for example, why these knowledge adapters are most useful for "normocytic anemia." We will try to fix these issues in future work.

Conclusions

This study serves as a preliminary exploration of capturing physician reasoning. By predicting patients' diagnoses based on physician comments, we aim to explore physician knowledge and the way they make judgments about the patients. We propose to inject domain knowledge from disparate sources into PLMs through adapters under the DAKI framework, enhancing their representation capability on clinical text. The experimental results demonstrate that capturing physician knowledge is feasible through the encoding of clinical text using PLMs, the representation capability and interpretability of which can be further improved when equipped with external domain knowledge. Notably, the transferability of the knowledge adapters, exemplified by comparable performance between DAKI-BERT and ClinicalBERT without access to extensive clinical corpora, underscores the potential for scalable and versatile applications across various institutional contexts and knowledge domains.

Acknowledgments

Research reported in this publication was supported by the National Center for Advancing Translational Sciences of the National Institutes of Health (award U01TR002062). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. This study was approved by the Mayo Clinic Institutional Review Board (#20-001137) for human participants research.

Data Availability

The Mayo Clinic IRP data set used in this study is not publicly available due to privacy risks and is not available by request. The source code of DAKI is available in the GitHub repository [6].

Conflicts of Interest

Author HL is the Associate Editor of *JMIR AI*.

Multimedia Appendix 1

Activation levels of the adapters placed at different layers of BERT (left) and ClinicalBERT (right), respectively.

[PNG File, 46 KB - [ai_v3i1e56932_app1.png](#)]

Multimedia Appendix 2

Greatest impact observed among the least successful targets.

[PNG File, 162 KB - [ai_v3i1e56932_app2.png](#)]

References

1. Smith K, Megyesi B, Velupillai S, Kvist M. Professional language in Swedish clinical text: linguistic characterization and comparative studies. *Nord J Linguist* 2014;37(2):297-323. [doi: [10.1017/s0332586514000213](https://doi.org/10.1017/s0332586514000213)]
2. Ma X, Xu P, Wang Z, Nallapati R, Xiang B. Domain adaptation with BERT-based domain classification and data selection. 2019 Presented at: Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo); 2019; Hong Kong, China p. 76-83 URL: <https://aclanthology.org/D19-6109/>

3. Sushil M, Suster S, Daelemans W. Are we there yet? Exploring clinical domain knowledge of BERT models. 2021 Presented at: Proceedings of the 20th Workshop on Biomedical Language Processing; 2021; Pennsylvania, United States p. 41-53. [doi: [10.18653/v1/2021.bionlp-1.5](https://doi.org/10.18653/v1/2021.bionlp-1.5)]
4. Alsentzer E, Murphy J, Boag W, Weng W, Jindi D, Naumann T, et al. Publicly available clinical BERT embeddings. 2019 Presented at: Proceedings of the 2nd Clinical Natural Language Processing Workshop; 2019; Minneapolis, Minnesota, USA p. 72-78.
5. Donnelly K. SNOMED-CT: The advanced terminology and coding system for eHealth. *Stud Health Technol Inform* 2006;121:279-290. [Medline: [17095826](https://pubmed.ncbi.nlm.nih.gov/17095826/)]
6. Lu Q, Dou D, Nguyen TH. Parameter-efficient domain knowledge integration from multiple sources for biomedical pre-trained language models. In: Findings of the Association for Computational Linguistics: EMNLP 2021. 2021 Presented at: The 2021 Conference on Empirical Methods in Natural Language Processing; 2021; Punta Cana, Dominican Republic p. 3855-3865. [doi: [10.18653/v1/2021.findings-emnlp.325](https://doi.org/10.18653/v1/2021.findings-emnlp.325)]
7. Wang X, Gao T, Zhu Z, Zhang Z, Liu Z, Li J, et al. KEPLER: A unified model for knowledge embedding and pre-trained language representation. *Transactions of the Association for Computational Linguistics* 2021:176-194. [doi: [10.1162/tacl_a_00360](https://doi.org/10.1162/tacl_a_00360)]
8. Franz L, Shrestha YR, Paudel B. A deep learning pipeline for patient diagnosis prediction using electronic health records. In: ArXiv. 2020 Presented at: The 19th International Workshop on Data Mining in Bioinformatics; 2020; San Diego, CA, USA URL: <https://arxiv.org/abs/2006.16926>
9. Johnson AE, Pollard TJ, Shen L, Lehman LH, Feng M, Ghassemi M, et al. MIMIC-III, a freely accessible critical care database. *Sci Data* 2016;3:160035 [FREE Full text] [doi: [10.1038/sdata.2016.35](https://doi.org/10.1038/sdata.2016.35)] [Medline: [27219127](https://pubmed.ncbi.nlm.nih.gov/27219127/)]
10. O'Shea K, Nash R. An introduction to convolutional neural networks. ArXiv. 2015. URL: <https://arxiv.org/abs/1511.08458> [accessed 2015-11-26]
11. Zoabi Y, Deri-Rozov S, Shomron N. Machine learning-based prediction of COVID-19 diagnosis based on symptoms. *NPJ Digit Med* 2021;4(1):3 [FREE Full text] [doi: [10.1038/s41746-020-00372-6](https://doi.org/10.1038/s41746-020-00372-6)] [Medline: [33398013](https://pubmed.ncbi.nlm.nih.gov/33398013/)]
12. Huang K, Altsaar J, Ranganath R. ClinicalBERT: modeling clinical notes and predicting hospital readmission. ArXiv. 2019. URL: <https://arxiv.org/abs/1904.05342> [accessed 2019-04-10]
13. Wang Z, Wen R, Chen X, Cao S, Huang S, Qian B, et al. Online disease diagnosis with inductive heterogeneous graph convolutional networks. 2021 Presented at: Proceedings of the Web Conference; 2021; New York, NY, United States p. 3349-3358. [doi: [10.1145/3442381.3449795](https://doi.org/10.1145/3442381.3449795)]
14. Choi E, Xu Z, Li Y, Dusenberry M, Flores G, Xue E, et al. Learning the graphical structure of electronic health records with graph convolutional transformer. 2020 Presented at: Proceedings of the AAAI Conference on Artificial Intelligence; 2020; Washington, DC, USA p. 606-613 URL: <https://ojs.aaai.org/index.php/AAAI/article/view/5400> [doi: <https://doi.org/10.1609/aaai.v34i01.5400>]
15. Hosseini A, Chen T, Wu W, Sun Y, Sarrafzadeh M. HeteroMed: Heterogeneous information network for medical diagnosis. 2018 Presented at: Proceedings of the 27th ACM International Conference on Information and Knowledge Management; 2018; New York, NY, United States p. 763-772 URL: <https://dl.acm.org/doi/10.1145/3269206.3271805> [doi: <https://doi.org/10.1145/3269206.3271805>]
16. Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* 2020;36(4):1234-1240 [FREE Full text] [doi: [10.1093/bioinformatics/btz682](https://doi.org/10.1093/bioinformatics/btz682)] [Medline: [31501885](https://pubmed.ncbi.nlm.nih.gov/31501885/)]
17. Gu Y, Tinn R, Cheng H, Lucas M, Usuyama N, Liu X, et al. Domain-specific language model pretraining for biomedical natural language processing. *ACM Trans Comput Healthcare* 2021;3(1):1-23. [doi: [10.1145/3458754](https://doi.org/10.1145/3458754)]
18. Lu Q, Dou D, Nguyen T. ClinicalT5: A generative language model for clinical text. In: Findings of the Association for Computational Linguistics: EMNLP 2022. 2022 Presented at: The 2022 Conference on Empirical Methods in Natural Language Processing; 2022; Abu Dhabi p. 5436-5443. [doi: [10.18653/v1/2022.findings-emnlp.398](https://doi.org/10.18653/v1/2022.findings-emnlp.398)]
19. Zhang Z, Han X, Liu Z, Jiang X, Sun M, Liu Q. ERNIE: Enhanced Language Representation with Informative Entities. 2019 Presented at: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics; 2019; Pennsylvania, United States p. 1441-1451 URL: <https://aclanthology.org/P19-1139/>
20. Sun T, Shao Y, Qiu X, Guo Q, Hu Y, Huang X, et al. CoLAKE: Contextualized language and knowledge embedding. 2020 Presented at: Proceedings of the 28th International Conference on Computational Linguistics; 2020; Barcelona, Spain p. 3660-3670 URL: <https://aclanthology.org/2020.coling-main.327/>
21. He Y, Zhu Z, Zhang Y, Chen Q, Caverlee J. Infusing disease knowledge into bert for health question answering, medical inference and disease name recognition. 2020 Presented at: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP); 2020; Pennsylvania, United States p. 4604-4614 URL: <https://aclanthology.org/2020.emnlp-main.372/> [doi: [10.18653/v1/2020.emnlp-main.372](https://doi.org/10.18653/v1/2020.emnlp-main.372)]
22. Xie Q, Bishop JA, Tiwari P, Ananiadou S. Pre-trained language models with domain knowledge for biomedical extractive summarization. *Knowledge-Based Systems* 2022;252:109460. [doi: [10.1016/j.knosys.2022.109460](https://doi.org/10.1016/j.knosys.2022.109460)]

23. Zhang X, Bosselut A, Yasunaga M, Ren H, Liang P, Manning C, et al. Greaselm: Graph reasoning enhanced language models. 2022 Presented at: International Conference on Learning Representations 2021 Oct 6; 2022; Virtual URL: <https://arxiv.org/abs/2201.08860>
24. Pakhomov S, Weston SA, Jacobsen SJ, Chute CG, Meverden R, Roger VL. Electronic medical records for clinical research: application to the identification of heart failure. *Am J Manag Care* 2007;13(6 Part 1):281-281 [FREE Full text] [Medline: [17567225](https://pubmed.ncbi.nlm.nih.gov/17567225/)]
25. Neumann M, King D, Beltagy I, Ammar W. ScispaCy: Fast and robust models for biomedical natural language processing. 2019 Presented at: Proceedings of the 18th BioNLP Workshop and Shared Task; 2019; Florence, Italy p. 319-327 URL: <https://aclanthology.org/W19-5034/>
26. Lehman E, Lialin V, Legaspi KE, Sy AJ, Pile PT, Alberto NR, et al. Learning to ask like a physician. In: Proceedings of the 4th Clinical Natural Language Processing Workshop. 2022 Presented at: Proceedings of the 4th Clinical Natural Language Processing Workshop; 2022; Seattle, WA p. 74-86. [doi: [10.18653/v1/2022.clinicalnlp-1.8](https://doi.org/10.18653/v1/2022.clinicalnlp-1.8)]
27. Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). 2018 Presented at: The 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies; 2019; Minneapolis, Minnesota, USA p. 4171-4186. [doi: [10.18653/v1/N19-1423](https://doi.org/10.18653/v1/N19-1423)]
28. Lan Z, Chen M, Goodman S, Gimpel K, Sharma P, Soricut R. ALBERT: A lite BERT for self-supervised learning of language representations. In: International Conference on Learning Representations. 2019 Presented at: International Conference on Learning Representations; 2019; New Orleans, Louisiana, USA URL: <https://arxiv.org/abs/1909.11942>
29. Hu E, Wallis P, Allen-Zhu Z, Li Y, Wang S, Wang L, et al. LoRA: low-rank adaptation of large language models. In: International Conference on Learning Representations. 2021 Presented at: International Conference on Learning Representations 2021 Oct 6; 2021; Virtual URL: <https://arxiv.org/abs/2106.09685>
30. Lundberg SM, Lee S. A unified approach to interpreting model predictions. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. 2017 Presented at: Proceedings of the 31st International Conference on Neural Information Processing Systems; 2017; Red Hook, NY, United States p. 4768-4777 URL: <https://dl.acm.org/doi/pdf/10.5555/3295222.3295230>
31. Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Res* 2004;32(Database issue):D267-D270 [FREE Full text] [doi: [10.1093/nar/gkh061](https://doi.org/10.1093/nar/gkh061)] [Medline: [14681409](https://pubmed.ncbi.nlm.nih.gov/14681409/)]
32. Köhler S, Gargano M, Matentzoglou N, Carmody LC, Lewis-Smith D, Vasilevsky NA, et al. The human phenotype ontology in 2021. *Nucleic Acids Res* 2021;49(D1):D1207-D1217 [FREE Full text] [doi: [10.1093/nar/gkaa1043](https://doi.org/10.1093/nar/gkaa1043)] [Medline: [33264411](https://pubmed.ncbi.nlm.nih.gov/33264411/)]

Abbreviations

ClinicalBERT: clinical bidirectional encoder representations from transformer

CTRL: knowledge controller

CUI: concept unique identifiers

DAKI: Diverse Adapters for Knowledge Integration

DS: disease adapter

HPO: Human Phenotype Ontology

IRP: impression/report/plan

KG: knowledge graph adapter

LoRA: low-rank

MIMIC-III: Medical Information Mart for Intensive Care III

NLP: natural language processing

PHI: protected health information

PLM: pretrained language model

SG: semantic grouping adapter

SHAP: SHapley Additive exPlanations

SNOMED CT: Systematized Medical Nomenclature for Medicine–Clinical Terminology

UMLS: Unified Medical Language System

Edited by A Mavragani; submitted 30.01.24; peer-reviewed by N Hong, L Ferreira, S Goyal; comments to author 09.04.24; revised version received 21.04.24; accepted 29.04.24; published 06.08.24.

Please cite as:

Lu Q, Wen A, Nguyen T, Liu H

Enhancing Clinical Relevance of Pretrained Language Models Through Integration of External Knowledge: Case Study on Cardiovascular Diagnosis From Electronic Health Records

JMIR AI 2024;3:e56932

URL: <https://ai.jmir.org/2024/1/e56932>

doi: [10.2196/56932](https://doi.org/10.2196/56932)

PMID: [39106099](https://pubmed.ncbi.nlm.nih.gov/39106099/)

©Qiu hao Lu, Andrew Wen, Thien Nguyen, Hongfang Liu. Originally published in JMIR AI (<https://ai.jmir.org>), 06.08.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR AI, is properly cited. The complete bibliographic information, a link to the original publication on <https://www.ai.jmir.org/>, as well as this copyright and license information must be included.

Original Paper

Health Care Professionals' and Parents' Perspectives on the Use of AI for Pain Monitoring in the Neonatal Intensive Care Unit: Multisite Qualitative Study

Nicole Racine^{1*}, PhD; Cheryl Chow^{2*}, PhD; Lojain Hamwi², BA; Oana Bucsea², MA; Carol Cheng³, MSc; Hang Du⁴, MA; Lorenzo Fabrizi⁵, PhD; Sara Jasim², MA; Lesley Johannsson⁶, BScN; Laura Jones⁵, PhD; Maria Pureza Laudiano-Dray⁵, MRes; Judith Meek⁷, MBBS, PhD; Neelum Mistry⁵, MSc; Vibhuti Shah⁸, MSc, MD; Ian Stedman⁹, LLB, PhD; Xiaogang Wang⁴, PhD; Rebecca Pillai Riddell², PhD

¹School of Psychology, University of Ottawa, Children's Hospital of Eastern Ontario Research Institute, Ottawa, ON, Canada

²Department of Psychology, York University, Toronto, ON, Canada

³Department of Nursing, Mount Sinai Hospital, Toronto, ON, Canada

⁴Department of Mathematics and Statistics, York University, Toronto, ON, Canada

⁵Department of Neuroscience, Physiology and Pharmacology, University College London, London, United Kingdom

⁶Mount Sinai Hospital, Toronto, ON, Canada

⁷Neonatal Care Unit, University College London Hospitals, London, United Kingdom

⁸Department of Pediatrics, Mount Sinai Hospital, Toronto, ON, Canada

⁹School of Public Policy and Administration, York University, Toronto, ON, Canada

*these authors contributed equally

Corresponding Author:

Nicole Racine, PhD

School of Psychology, University of Ottawa

Children's Hospital of Eastern Ontario Research Institute

136 Jean-Jacques Lussier

Ottawa, ON

Canada

Phone: 1 403 992 7869

Email: nracine2@uottawa.ca

Abstract

Background: The use of artificial intelligence (AI) for pain assessment has the potential to address historical challenges in infant pain assessment. There is a dearth of information on the perceived benefits and barriers to the implementation of AI for neonatal pain monitoring in the neonatal intensive care unit (NICU) from the perspective of health care professionals (HCPs) and parents. This qualitative analysis provides novel data obtained from 2 large tertiary care hospitals in Canada and the United Kingdom.

Objective: The aim of the study is to explore the perspectives of HCPs and parents regarding the use of AI for pain assessment in the NICU.

Methods: In total, 20 HCPs and 20 parents of preterm infants were recruited and consented to participate from February 2020 to October 2022 in interviews asking about AI use for pain assessment in the NICU, potential benefits of the technology, and potential barriers to use.

Results: The 40 participants included 20 HCPs (17 women and 3 men) with an average of 19.4 (SD 10.69) years of experience in the NICU and 20 parents (mean age 34.4, SD 5.42 years) of preterm infants who were on average 43 (SD 30.34) days old. Six themes from the perspective of HCPs were identified: regular use of technology in the NICU, concerns with regard to AI integration, the potential to improve patient care, requirements for implementation, AI as a tool for pain assessment, and ethical considerations. Seven parent themes included the potential for improved care, increased parental distress, support for parents regarding AI, the impact on parent engagement, the importance of human care, requirements for integration, and the desire for choice in its use. A consistent theme was the importance of AI as a tool to inform clinical decision-making and not replace it.

Conclusions: HCPs and parents expressed generally positive sentiments about the potential use of AI for pain assessment in the NICU, with HCPs highlighting important ethical considerations. This study identifies critical methodological and ethical perspectives from key stakeholders that should be noted by any team considering the creation and implementation of AI for pain monitoring in the NICU.

(*JMIR AI 2024;3:e51535*) doi:[10.2196/51535](https://doi.org/10.2196/51535)

KEYWORDS

pain monitoring; pain management; preterm infant; neonate; pain; infant; infants; neonates; newborn; newborns; neonatal; baby; babies; pediatric; pediatrics; preterm; premature; assessment; intensive care; NICU; neonatal intensive care unit; HCP; health care professional; health care professionals; experience; experiences; attitude; attitudes; opinion; perception; perceptions; perspective; perspectives; acceptance; adoption; willingness; artificial intelligence; AI; digital health; health technology; health technologies; interview; interviews; parent; parents

Introduction

Globally, an estimated 13.4 million babies were born preterm in 2020, accounting for about 1 in 10 of all babies born [1]. Unfortunately, a significant proportion of preterm infants require neonatal intensive care unit (NICU) due to their vulnerability to complications and health issues [2]. As part of their lifesaving care, preterm infants undergo an average of 10 to 16 painful procedures per day [3]. Unmanaged NICU pain has significant developmental consequences [4,5] and is one of the largest sources of severe emotional distress in parents [6]. Pain assessment and management is a critical aspect of care in the NICU [7]. Traditional pain assessment methods in the NICU rely on observational tools [8,9]. However, there are several challenges with these methods, including bias and subjectivity, staff time resources, and potential variability in interpretation [10-12]. Given these challenges, innovative approaches are needed to improve existing pain assessment practices. Artificial intelligence (AI), which includes machine learning (ie, using a machine to extract knowledge from data and learn autonomously), is one technology that has shown tremendous potential in the health care field, and this potential may also inform the development of clinical decision support systems [13]. Specifically, AI-based technology can analyze large volumes of behavioral, physiological, and brain imaging data to provide suggestions with regard to infant pain assessment at the point of care.

Current evidence about the use of AI in the assessment and monitoring of infant pain appears to be promising [14,15]. Preliminary algorithms to monitor vital signs [16], such as heart rate, respiratory rate, and oxygen saturation, of preterm infants have been developed, all of which provide physiological indications of pain or distress as well as systems that incorporate behavioral indicators (eg, face movements, body movements, and crying) to predict pain [17]. Although there is immense potential for these new technologies to revolutionize how neonatal pain is assessed and monitored in the NICU, a limited understanding of the perspectives of key stakeholders with regard to this emerging technology exists, that is, health care professionals (HCPs) and parents. These perspectives are essential for the successful implementation of this technology in clinical practice.

Studies exploring the attitudes and trust of clinicians toward AI in health care found that while there is recognition of AI's

potential benefits, concerns persist about reliability, transparency, data privacy, potential loss of autonomy in decision-making, and potential misinterpretation [18-21]. Factors such as age, education level, and previous experience with AI influenced attitudes and trust in AI technologies [21].

There is a growing interest in the application of AI technologies in health care, particularly in neonatal and pediatric care [14]. However, little is known about the perspectives of HCPs and parents on the use of AI for pain assessment in the NICU. Pain is a significantly different context warranting focused study because infants cannot verbalize for themselves. This study explores the perspectives of health care professionals and parents with regard to automated pain assessment using AI technology in the NICU. This study will inform the implementation of AI, specifically machine learning technology in the NICU, leading to more effective pain assessment and management strategies.

Methods

Ethical Considerations

Ethics approval for this qualitative study was granted from all study sites, including York University (2020-034), Mount Sinai Hospital (MSH; 19-0252-A), and University College London Hospital (UCLH; 11/LO/0350). Informed consent was obtained from all participants. All data were deidentified. Individuals were provided with a CAD \$10 (approximately US \$7) gift card to a local coffee shop for their participation.

Setting and Design

Data collection occurred at 2 tertiary care NICUs: MSH (Toronto, Canada) and UCLH (London, United Kingdom). The study is part of a larger project focused on the use of AI, specifically the development of a machine learning algorithm, to assess infant pain in the NICU. Participants consisted of 20 HCPs (nurses, physicians, and allied health professionals) and 20 parents (mothers and fathers). Recruitment at MSH took place from February to March 2020, and recruitment at UCLH took place from July 2021 to October 2022. Interviews at MSH occurred in person at the hospital, whereas interviews at UCLH were web-based and conducted using a secure Zoom platform (Zoom Video Communications). This difference was due to the onset of the COVID-19 pandemic after the study had launched, which delayed the UK interviews and necessitated the use of a secure web platform. For HCPs, eligibility criteria were (1) currently providing care to infants at one of the NICUs and (2)

trained as either a nurse, physician, or other health professionals (ie, outreach staff and consultant practice educator). For parents, eligibility criteria included being 18 years and older of age, having an infant who was currently receiving care in the NICU, and being fluent in English, orally (in order to respond to complex questions in the interview). Using a purposive sampling approach, all participants were initially approached by 1 clinical member of staff on the unit and asked if they were interested in participating in the study. Only families where the parent was at least 18 years of age and spoke English were approached. If interested, they received additional information, and a time was scheduled for an interview.

Following introductions and the completion of the consent form, 30-minute semistructured interviews were conducted by a member of the research team (NR, C Chow, and L Johannsson) in a private clinic room (MSH) or web-based room (UCLH). Baseline demographic information was collected at the outset of the meeting followed by a series of questions (10 for HCPs and 9 for parents) pertaining to the use of AI to inform NICU decision-making related to the assessment of infant pain. Notes were taken during the interviews to supplement transcripts. Interviewers read an initial script providing a definition of AI and providing context for the study. In-person interviews were recorded using a digital audio recorder, whereas web-based interviews were recorded using privacy-compliant web software (Zoom) and stored on a secure server. All participants were debriefed following the interview and provided with a gift card to a local coffee shop as a token of appreciation. Standards for Reporting Qualitative Research were followed for this study ([Multimedia Appendix 1](#) [22]).

Development of the Interview Guides

Using a grounded theory approach [23], the goal of the qualitative interviews was to generate detailed knowledge about HCPs' and parents' understandings and perceptions of the use of AI in the NICU to assist with infant pain assessment and management. Specifically, we sought to gain insight into HCPs' and parents' understanding of AI, perceived implications of this technology, potential benefits of the technology, and barriers to its use in the NICU setting. Two interview guides were developed to address the diverse perspectives of HCPs ([Multimedia Appendix 2](#)) and parents ([Multimedia Appendix 3](#)). The interview guides were developed collaboratively by members of the research team (RPR and NR), who are clinical psychologists with previous experience in conducting qualitative research with both HCPs and parents in the NICU and other pediatric medical settings [24,25]. The guides were reviewed and edited based on the feedback from team members with NICU clinical expertise (VS, C Chow, JM, and MPL-D) as well as ethical or legal or social expertise related to AI (IS). Interviews were conducted by 2 postdoctoral fellows (NR and C Chow) and 1 research staff (L Johannsson). A decision was made in advance to review and make necessary changes to the questions after the first interviews were conducted at each site based on participant comprehension and feedback. Based on the review, no major alterations were required. Participants had the opportunity to provide any additional comments or feedback at the end of the interview. Interviews were conducted until saturation was reached [26].

Data Processing and Analysis

The interview audio recordings were anonymized and transcribed by 1 research assistant and independently double-checked by members of the research team. Transcripts were subsequently analyzed using 6 phases of thematic analysis (ie, familiarization, generating codes, identifying themes, reviewing themes, naming themes, and report writing) [27]. Data analyses took place from February to April 2023. There were 3 analysis leads (NR, C Chow, and RPR) who took primary responsibility for developing the code book, overseeing the coding process, and developing themes based on the codes generated. As a first step, the analysis leads familiarized themselves with the data by reading and making notes on the transcripts. Responses were examined for differences between the 2 sites (eg, unique considerations related to the country, time, or modality via in-person vs web-based) or any effects that may have necessitated a different analysis pathway. It was determined that there were no differences, and we proceeded with analyzing the transcripts together. Next, a list of initial codes was generated independently by the analysis leads prior to a consensus meeting. Two consensus meetings were held, where all codes were reviewed and agreed upon. Subsequently, the analysis leads (NR, RPR, and C Chow) ran a 90-minute training session with 10 coders to familiarize them with the codes that have been created. All coders (LH, SJ, OB, VS, MPL-D, C Cheng, IS, HD, NM, and L Jones) were members of an interdisciplinary research team (ie, neurobiology, behavioral neuroscience, neurophysiology, psychology, medicine, nursing, and law) with research backgrounds in pediatric health care, with most specializing in infant care. Each transcript was coded twice. The average percent agreement (ie, the number of times 2 individuals agreed upon a code divided by the total number of units of observation that were rated) across transcripts between coders for the HCP and parent transcripts was 0.77, which is adequate [28]. Next, the analysis leads reviewed the coded transcripts and collated codes for each question. The analysis leads met and generated relevant potential themes and a thematic map based on the data. Finally, examples were selected to accompany each theme, which are presented in the results below. Summary statistics of all demographic variables were conducted in SPSS (version 28; IBM Corp).

Results

Participant Characteristics

The participant characteristics are shown in [Table 1](#). In total, 90% (n=18) of HCPs were university-educated and had extensive experience in the NICU (mean 19.4, SD 10.69 years; range 4-37 years). For HCPs, 55% (n=11) reported "Western" cultural heritages (eg, Canadian, British, and Australian), 5% (n=1) African, 15% (n=3) East Asian, 10% (n=2) Caribbean, 10% (n=2) South Asian, and 5% (n=1) not reported. For parents, 80% (n=16) reported "Western" cultural heritages (eg, Canadian, European, or Australian), 5% (n=1) Asian, 5% (n=1) Middle Eastern, and 10% (n=2) not reported. Most parents who participated across both sites were mothers (n=17, 85%) with a mean age of 34 (SD 5.42) years. In total, 90% (n=18) of parents had a university education or higher.

Table 1. Participant demographic characteristics.

Characteristics	Health care providers (n=10 each)		Parents (n=10 each)	
	Mount Sinai Hospital	University College London Hospital	Mount Sinai Hospital	University College London Hospital
Gender, n (%)				
Women	9 (90)	8 (80)	8 (80)	9 (90)
Men	1 (10)	2 (20)	2 (20)	1 (10)
Age (years), mean (SD)	— ^a	—	34.56 (6.04)	34.2 (5.14)
Postnatal age of infant (days), mean (SD)	—	—	28.11 (24.09)	57.5 (29.52)
Highest level of education, n (%)				
Graduate school or professional training	6 (60)	7 (70)	3 (30)	9 (90)
University graduate	2 (20)	3 (30)	5 (50)	1 (10)
Partial university	0 (0)	0 (0)	0 (0)	0 (0)
Trade school or community college	1 (10)	0 (0)	1 (10)	0 (0)
High school graduate	0 (0)	0 (0)	0 (0)	0 (0)
Less than high school	0 (0)	0 (0)	1 (10)	0 (0)
Not reported	1 (10)	0 (0)	0 (0)	0 (0)
Heritage culture, n (%)				
African	1 (10)	0 (0)	0 (0)	0 (0)
Asian	2 (20)	1 (10)	1 (10)	0 (0)
Australia or New Zealand	0 (0)	0 (0)	0 (0)	2 (20)
Caribbean	2 (20)	0 (0)	0 (0)	0 (0)
Canadian	1 (10)	0 (0)	5 (50)	0 (0)
European	3 (30)	7 (70)	1 (10)	8 (80)
Middle Eastern	0 (0)	0 (0)	1 (10)	0 (0)
South Asian	0 (0)	2 (20)	0 (0)	0 (0)
Not reported	1 (10)	0 (0)	2 (20)	0 (0)
Type of health care professional, n (%)				
Physician	5 (50)	3 (30)	—	—
Registered nurse	5 (50)	4 (40)	—	—
Other health professional	0 (0)	3 (30)	—	—
Experience (years), mean (SD)	22 (8.55)	16 (12.18)	—	—

^aNot available.

HCP Themes

Six themes emerged from the thematic analysis on the HCP interviews. Each theme, a description, and representative quotes are presented in [Table 2](#). HCP themes and subthemes are presented in [Figure 1](#). First, in the context of their comfort with incorporating new AI technology, HCPs reported limited experience with AI technology in the NICU (1 HCP was part of a research study at another institution), and they were comfortable using other forms of technology. Second, HCPs identified some concerns with regard to the integration of AI for pain assessment in the NICU. Some of these concerns included increased distress from knowing clinicians were inflicting pain and extra workload for HCPs, increased stress

for parents, and decreased opportunities for parent-child bonding, as well as fears related to overreliance on AI technology and the overuse of medication to manage pain. Despite these concerns, the third theme emerged surrounding several benefits that AI could bring to the NICU context. Notably, HCPs identified increased awareness of infant pain, early detection and diagnosis of clinical changes, increased efficiency, and standardization of pain assessment, as well as the potential to inform the development of better pain management strategies. From a practical standpoint, the fourth theme identified requirements to facilitate the implementation of AI in the NICU, including the size of machinery, staff training, as well as clearly communicating the validity, sensitivity, and specificity of the algorithm being used. The fifth

theme that was unanimously shared was the idea that using AI for pain assessment in the NICU would be a tool for HCPs to use but could not replace the clinical judgment and decision-making of an HCP. Concerns related to how the next generation of HCPs would be trained to ensure that they have both the clinical and technological skills to operate in the NICU were described, given the potential overreliance on technology. Finally, HCPs identified the potential for ethical concerns related to an AI algorithm for constant pain monitoring in the NICU,

specifically, issues related to the disagreement between HCP and the AI algorithm, implications of pain monitoring in the absence of pain management, as well as the need to audit the algorithm. Overall, there was general acceptability for the benefits, use, and integration of AI technology for pain assessment in the NICU, with keen identification of the potential work-related, structural, technological, and ethical issues that would need to be addressed to facilitate implementation.

Figure 1. Themes and subthemes generated from qualitative interviews with HCPs on their perspectives about using AI to assess pain in the NICU. AI: artificial intelligence; Ax: assessment; HCP: health care professional; NICU: neonatal intensive care unit.

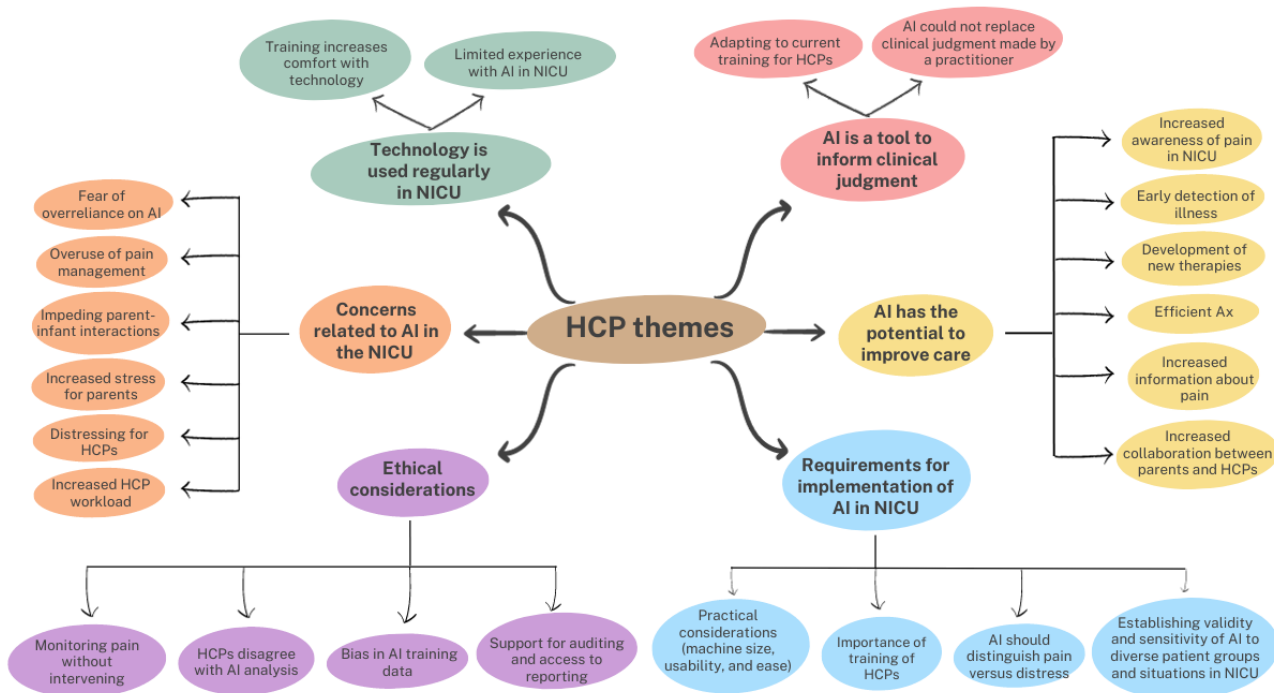


Table 2. Key themes identified by HCPs^a with regard to the use and integration of AI^b for pain assessment in the NICU^c.

Theme	Description	Representative quote
Technology is used regularly in the NICU	HCPs shared that despite having limited experience with AI specifically, they use technology to inform their clinical decision-making and they feel comfortable using the technology that is currently available.	<ul style="list-style-type: none"> “It informs everything. I think that’s one of the things that working in intensive care is that we use technology and monitoring to inform a lot of our decisions.”
Concerns of AI integration for pain assessment in the NICU	HCPs identified concerns related to the integration of AI in the NICU. It specifically increased the workload for HCPs and increased distress, knowing they were potentially inflicting pain on an infant. They also reported that constant pain monitoring could increase stress for parents and that added machinery could inhibit parent-child bonding. Concerns were also identified with regard to the overreliance on what the algorithm reported and the overuse of pain pharmaceuticals to manage pain.	<ul style="list-style-type: none"> Increased HCP distress: “I’m not sure cause you imagine like how upsetting it would be like you know I’m doing a diaper change and this thing is telling me the baby is in pain.” Increased workload: “I think there would be some negative feedback towards having extra work to be done.” Fear of overreliance on the AI: “The disadvantages would be that we become over reliant on it. And just because the machine says the baby’s not in pain, then it could be dismissed as the baby isn’t in pain, when actually if you look at the baby, you can tell they’re in pain.” Increased parent stress: “It can cause stress ... Unnecessary stress.” Impeding parent-child bonding: “I can see it taking away from looking at babies...you see parents, particularly looking at their monitor alarms, for whatever reason, they look more at the monitor than actually what their baby’s doing.”
AI has the potential to improve pain assessment and management	HCPs indicated there are several ways in which integrating constant pain monitoring in the NICU could improve clinical care, including the development of new therapies, early diagnosis of difficulties, detection of changes in clinical presentation, increased awareness of infant pain, increased efficiency of pain assessment, increased standardization of pain assessment, and increased collaboration between HCPs and parents.	<ul style="list-style-type: none"> “I think it’s good that um there is a form of technology that can give us more information about pain in this population because I think there’s a lot of unknown and I think well I know for myself like I said I can’t honestly say that I’m always thinking about if this baby is in pain or what kind of pain this baby is in when doing a procedure.” “I think it would give them more time to obviously focus on other aspects of their work instead of having to score every half an hour or so to proceed and enter the data as it is at the moment.”
Requirements for implementation of AI in NICU	HCPs described structural (ie, machine size and invasiveness of machinery) requirements for implementing AI in the NICU. Specifically, machinery would need to be small and noninvasive. HCPs indicated that training staff to understand and interpret the output provided by the technology is important. They also indicated that the algorithm would need to be properly validated and sensitive for detecting pain in diverse patient groups and situations.	<ul style="list-style-type: none"> Structural requirements: “It depends how invasive the technology is. When you have a 450 gram baby in front of you. Even putting on things like more monitors actually occludes your that visual assessment of the child. So I think there can be barriers.” Importance of training: “I think obviously, it’s all about training ... everybody understands how it works and the benefits.”
AI is a tool to inform clinical pain assessment and management	HCPs indicated that AI in the NICU should be viewed as a tool to inform clinical decision-making but not as a replacement. They also indicated that the integration of this technology would have implications for the training of new HCPs to ensure they have the ability to understand how this tool could inform their own clinical assessment.	<ul style="list-style-type: none"> “I like using technology but as long as it doesn’t replace my ability to provide comfort and care” “If I’m gonna make it’s just detection of pain, I think it’d be fairly comfortable with that. Because then I can react to that. Whereas if it’s making medical decision on the treatment, a baby’s receiving, I think that will be a completely different scenario.”
Ethical concerns with constant pain monitoring may occur	HCP indicated the need to be aware of ethical concerns like the potential bias in AI algorithms, disagreements between HCPs and the AI’s output, and the implications of constant pain monitoring without intervening. HCPs also indicated that algorithms would need to be audited and monitored over time.	<ul style="list-style-type: none"> “And then you have to decide, what you want to do about it. And then you have to decide, in a medical-legal issue whether to believe A.I. or the clinician and that will be interesting.”

^aHCP: health care professional.^bAI: artificial intelligence.^cNICU: neonatal intensive care unit.

Parent Themes

Seven overarching themes were identified with parents (Table 3). Parent themes and subthemes are presented in Figure 2. First, parents indicated it would be desirable to know if their infants were in pain because there are limited ways of assessing neonatal pain and it would provide useful information to HCPs to improve their infant’s care. However, the second theme arose about the emotional toll that may be experienced by parents. Some parents noted heightened distress from knowing their infant was experiencing pain. The third theme revolved around a preference to have parents decide for themselves whether they wanted continuous pain monitoring using AI. The fourth theme was that parents indicated wanting support to interpret and understand the constant pain monitoring. That is, they would want HCPs to explain their decision-making process as well as how the pain assessment provided by the AI was being used.

The fifth theme was that parents perceived their current level of engagement in their infant’s care to be quite high and they did not think constant pain monitoring would change this engagement. The sixth theme was that most parents would not trust an AI to make an independent decision about their infant’s pain but rather believe it should be incorporated as a tool by HCPs to make a clinical decision. Parents voiced that there would be potential for error in the AI’s assessment and that verification by an HCP would be important. Finally, parents identified requirements related to AI integration in the NICU. Specifically, they are concerned about privacy since large amounts of data would be collected and therefore would need to be kept secure. They also identified that the algorithm should be developed in a nonbiased way and that generalizability of the algorithm across infant presentations and contexts would be needed.

Figure 2. Themes and subthemes generated from qualitative interviews with parents on their perspectives of using AI to assess pain in the neonatal intensive care unit. AI: artificial intelligence; HCP: health care professional.

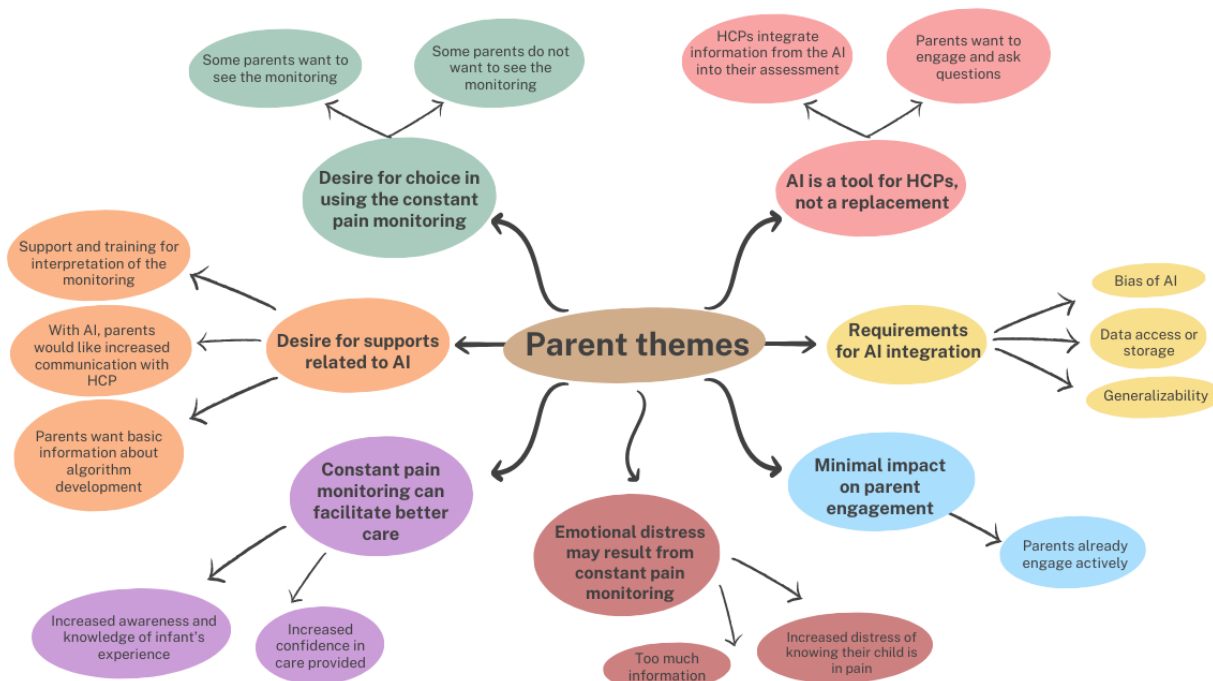


Table 3. Key themes identified by parents with regard to the use and integration of AI^a for pain assessment in the NICU^b.

Theme	Description	Representative quote
Constant pain monitoring can facilitate better care	Parents indicated there are advantages to constant pain monitoring (eg, increase in awareness of infant's experience and confidence in care provided).	<ul style="list-style-type: none"> • “But then it could also help the parent, could help us understand the baby a bit more and maybe bond maybe a bit more or communicate in a way with the baby more.”
Emotional distress may result from constant pain monitoring	Parents shared disadvantages to constant pain monitoring, such as too much information or distress associated with knowing their child is in pain.	<ul style="list-style-type: none"> • “My gut is saying, as a parent, well, of course. But I'm wondering whether you can have almost too much information, where if certain things, I definitely would be in this position, where if certain things had to be done to my child, life and death or even just less serious, but they needed to be done for, you know, health reasons, how productive is it for a parent to know exactly how much pain their child is in.”
Desire for choice in using the constant pain monitoring	Parents indicated that they would like to be given a choice to view the constant pain monitoring.	<ul style="list-style-type: none"> • “You should have a choice in the same way as like, you can choose to look at lots of the information about your baby or not.” • “I mean I would want to know if my baby is in pain or not. But maybe some parents are ok with or don't want to know about their baby's pain but to me I would definitely want to see.”
Desire for support related to AI	Parents indicated that they would want communication from staff and support to understand and interpret the constant pain monitoring. They would also like basic information about how the algorithm was developed and makes its predictions.	<ul style="list-style-type: none"> • “Because even now I don't want to do anything unless the nurse is there ... but you see that number go up as you're as you're caring for the baby you might be or I might be a little apprehensive um but with the reassurance of the nurse or if you can see that once the baby is settled down the baby is more comfortable again then you know that it's ok.” • “I would want to know, and I would want it to be very clear why those decisions were made. I would want, if we were using kind of artificial intelligence, what kind of almost a report on why those decisions were made and why it was recommended that XYZ happened as a result.”
Minimal impact on parent engagement	Parents indicated that constant pain monitoring would minimally impact their level of active engagement in the newborn's care as most reported that they already engaged at a high level.	<ul style="list-style-type: none"> • “You know I'm not sure that it would change how engaged I would be because I think you know you can use other metrics as like surrogate of pain as well and being at the bedside you can still be engaged in her care but I guess it could be interested to ask you know like when we should up like you know how were her pain scores overnight or something like that. And you know get that data and get that information from the bedside nurse. But I don't think it would dramatically change the engagement.”
AI for pain monitoring is a tool for HCPCs ^c , not a replacement	Parents indicated that constant pain monitoring should be used as a tool to inform clinical judgment.	<ul style="list-style-type: none"> • “Yeah no I would like the doctor so I could also ask questions and you know it's yeah a tool to assess or to inform them” • “And I think it makes sense that the physician in the bedside needs to integrate that with what their clinical assessment is” • “It would be a good thing if doctors were checking in to validate that the AI was right and if they disagree they should definitely question it [...] maybe the model is wrong or like maybe the model just needs to be tweaked and it needs doctors and scientists to question it right? It's probably a good thing.”

Theme	Description	Representative quote
Requirements for AI integration in the NICU	Parents indicated that it would be important to consider how data might be collected and used by the AI, how to reduce bias in the development of the algorithm, and how to ensure that the algorithm was generalizable across infants and contexts.	<ul style="list-style-type: none"> “... questions about the data it was collecting and where that was going and who’s using that data. So obviously, the monitoring there’s a lot of information there.” “I would be concerned if a model was created that the way in which it was created was maybe not ethical but I’m I know there’s all kinds of laws and things like that but I was just thinking about how that might work.” “And then the sample size and the how many different like every baby is different and every baby’s pain tolerance is different how do you know that you’ve got all your bases covered for all the different scenarios.”

^aAI: artificial intelligence.

^bNICU: neonatal intensive care unit.

^cHCP: health care professional.

Discussion

Principal Findings

This international study includes the perspectives of both HCPs (ie, physicians and nurses) and parents regarding the use of AI technology in the NICU setting. These perspectives offer critical insights to help inform the development of potential AI technology on infant pain management and integration of this technology as part of clinical decision support systems. We found that both HCPs and parents were supportive of the use of AI technology in predicting infant pain. Both HCPs and parents recognized that AI has the potential to improve care in the NICU setting. Other studies have also identified similar benefits including earlier detection of illness, increased collaboration and communication, and development of new treatments that further support the use of AI in clinical settings [29,30].

In line with previous research [31], this study also found that HCPs and parents had similar concerns on the use of AI technologies in the NICU setting, including effectiveness and accuracy, fear of overreliance, and shared decision-making over the use of AI technology. Furthermore, we identified additional themes from the perspectives of parents regarding the importance of receiving support for interpreting and understanding constant pain monitoring. Interestingly, most parents indicated that they would prefer the choice to have access to constant pain monitoring in real time, as it could impact parents differently. Moreover, both HCPs and parents identified the importance of using AI as an adjunctive tool to inform clinical decisions. That is, both parents and HCPs seemed in favor of using AI to augment human intelligence and support more informed clinical decision-making [32] rather than automating any aspect of clinical care. Similar to youth and adult patients, parents of infants in the NICU were concerned about the risk of clinician replacements and emphasized the importance of the human element (ie, HCP’s presence at the bedside) in clinical care [30,33,34]. Clinicians also warned about the potential for diminished skills and overreliance on technology for the next generation of clinicians with regard to

pain assessment at the bedside. It is worth noting that clinical decision-making and responsibility continue to rest with clinicians, and there is currently no legislation that would allow automated health care decisions by an AI [35]. These new emerging themes could potentially help inform the future development of AI tools in the NICU setting as well as the training of future HCPs working in the NICU. Findings from this study could be used to justify increased training, engagement, and consultation with health care professionals as AI is implemented in the NICU.

Interestingly, we found very similar responses and results across countries as well as interview modalities. This is not surprising as both the United Kingdom and Canada follow similar protocols within the NICUs as both have public health care systems. Additionally, structured interviews, such as those conducted in this study, work equally well in face-to-face or web-based studies [36]. Furthermore, the interviewers were the same across both contexts. We also found that both HCPs and parents had limited experience with the use of AI in the NICU, meaning that all the responses garnered in this study were hypothetical in nature. Had participants had exposure, they may have provided different responses with regard to the feasibility and use of this technology. Future research prior to and during the implementation process will be important to capture these perspectives.

Limitations

There are some limitations to this study that should be considered when interpreting our results. First, interviews were conducted with HCPs at 2 large, tertiary-care, academic hospitals in Canada and the United Kingdom that are at the forefront of technological advancement in the NICU. As such, the perspectives of HCPs in this study may not be generalizable to smaller, less well-resourced care settings. Second, parents included in this study were highly educated, which may limit generalizability to parents with lower educational attainment, which is also a known risk factor for preterm birth [37]. Moreover, parents were recruited into the sample if they spoke English, which may have resulted in a less culturally diverse sample. Third, many of the themes that were identified by HCPs

and caregivers were broad in that they were not referring to the use of AI specifically but rather the use of clinical decision support systems (ie, a clinician using technology like AI to help inform their decisions related to care). As both technology and terminology evolve in the medical context, it will be important to disentangle opinions related to the technology itself as opposed to its use as a clinical decision-making tool. Finally, questions asked of HCPs and parents differed with more emphasis placed on general technology with HCPs and on neonatal pain for parents. This may have had an impact on the responses that were generated. As AI-related technology is integrated into medical settings, future qualitative research may focus specifically on pain-related questions.

Conclusions

Based on detailed interviews with 40 HCPs and parents across 2 large NICUs in publicly funded hospitals in Canada and the United Kingdom, our overall findings indicate that both HCPs and parents view the integration of an AI algorithm for constant pain monitoring to have potential benefits and to be an acceptable practice. Notably, HCPs identified several ways in which constant pain monitoring could improve the clinical care provided in the NICU. Both HCPs and parents were balanced in their perspectives and identified potential disadvantages as well as requirements for the successful implementation of an AI tool for pain assessment. Taken together, there is immense promise as well as major structural, ethical, and methodological considerations for the development and implementation of AI technology in the NICU setting.

Acknowledgments

This project was funded by the Canadian Institutes of Health Research, Natural Sciences and Engineering Research Council, Social Sciences and Humanities Research Council, and Collaborative Health Research Projects (principal investigator: RPR; grant 1001-2019-0004). LF, JM, L Jones, and MPL-D were funded by the Medical Research Council UK (grant MR/S003207/1). The funders had no role in data collection, interpretation, and reporting. NR receives funding through the Chair in Child and Youth Mental Health at the Children's Hospital of Eastern Ontario and the University of Ottawa.

Authors' Contributions

RPR, NR, and C Chow conceptualized the paper, developed the qualitative coding system, conducted the qualitative interviews, did the data analysis, and wrote the first draft of the paper. LH contributed to writing the introduction. L Johannsson and C Cheng contributed to data collection. LH, OB, C Cheng, HD, LF, SJ, L Johannsson, L Jones, MPL-D, JM, NM, VS, IS, and XW contributed to data coding. All authors reviewed the final paper. RPR was responsible for obtaining the primary funding.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Standards of Reporting Qualitative Research [22].

[[DOCX File , 19 KB - ai_v3i1e51535_app1.docx](#)]

Multimedia Appendix 2

Interview guide for health care professionals.

[[DOCX File , 16 KB - ai_v3i1e51535_app2.docx](#)]

Multimedia Appendix 3

Interview guide for caregivers.

[[DOCX File , 16 KB - ai_v3i1e51535_app3.docx](#)]

References

1. Ohuma EO, Moller AB, Bradley E, Chakwera S, Hussain-Alkhateeb L, Lewin A, et al. National, regional, and global estimates of preterm birth in 2020, with trends from 2010: a systematic analysis. *Lancet* 2023;402(10409):1261-1271 [FREE Full text] [doi: [10.1016/S0140-6736\(23\)00878-4](https://doi.org/10.1016/S0140-6736(23)00878-4)] [Medline: [37805217](https://pubmed.ncbi.nlm.nih.gov/37805217/)]
2. Blencowe H, Cousens S, Oestergaard MZ, Chou D, Moller AB, Narwal R, et al. National, regional, and worldwide estimates of preterm birth rates in the year 2010 with time trends since 1990 for selected countries: a systematic analysis and implications. *Lancet* 2012;379(9832):2162-2172 [FREE Full text] [doi: [10.1016/S0140-6736\(12\)60820-4](https://doi.org/10.1016/S0140-6736(12)60820-4)] [Medline: [22682464](https://pubmed.ncbi.nlm.nih.gov/22682464/)]
3. Johnston C, Barrington KJ, Taddio A, Carbajal R, Filion F. Pain in Canadian NICUs: have we improved over the past 12 years? *Clin J Pain* 2011;27(3):225-232. [doi: [10.1097/AJP.0b013e3181fe14cf](https://doi.org/10.1097/AJP.0b013e3181fe14cf)] [Medline: [21178602](https://pubmed.ncbi.nlm.nih.gov/21178602/)]

4. Grunau RE, Holsti L, Peters JWB. Long-term consequences of pain in human neonates. *Semin Fetal Neonatal Med* 2006;11(4):268-275. [doi: [10.1016/j.siny.2006.02.007](https://doi.org/10.1016/j.siny.2006.02.007)] [Medline: [16632415](https://pubmed.ncbi.nlm.nih.gov/16632415/)]
5. Woodward LJ, Moor S, Hood KM, Champion PR, Foster-Cohen S, Inder TE, et al. Very preterm children show impairments across multiple neurodevelopmental domains by age 4 years. *Arch Dis Child Fetal Neonatal Ed* 2009;94(5):F339-F344. [doi: [10.1136/adc.2008.146282](https://doi.org/10.1136/adc.2008.146282)] [Medline: [19307223](https://pubmed.ncbi.nlm.nih.gov/19307223/)]
6. Franck LS, Allen A, Cox S, Winter I. Parents' views about infant pain in neonatal intensive care. *Clin J Pain* 2005;21(2):133-139. [doi: [10.1097/00002508-200503000-00004](https://doi.org/10.1097/00002508-200503000-00004)] [Medline: [15722806](https://pubmed.ncbi.nlm.nih.gov/15722806/)]
7. Anand KJ, Scalzo FM. Can adverse neonatal experiences alter brain development and subsequent behavior? *Biol Neonate* 2000;77(2):69-82. [doi: [10.1159/000014197](https://doi.org/10.1159/000014197)] [Medline: [10657682](https://pubmed.ncbi.nlm.nih.gov/10657682/)]
8. Ballantyne M, Stevens B, McAllister M, Dionne K, Jack A. Validation of the premature infant pain profile in the clinical setting. *Clin J Pain* 1999;15(4):297-303. [doi: [10.1097/00002508-199912000-00006](https://doi.org/10.1097/00002508-199912000-00006)] [Medline: [10617258](https://pubmed.ncbi.nlm.nih.gov/10617258/)]
9. Merkel SI, Voepel-Lewis T, Shayevitz JR, Malviya S. The FLACC: a behavioral scale for scoring postoperative pain in young children. *Pediatr Nurs* 1997;23(3):293-297. [Medline: [9220806](https://pubmed.ncbi.nlm.nih.gov/9220806/)]
10. Riddell RP, Flora DB, Stevens S, Greenberg S, Garfield H. The role of infant pain behaviour in predicting parent pain ratings. *Pain Res Manag* 2014;19(5):e124-e132 [FREE Full text] [doi: [10.1155/2014/934831](https://doi.org/10.1155/2014/934831)] [Medline: [25299475](https://pubmed.ncbi.nlm.nih.gov/25299475/)]
11. Bellieni CV, Cordelli DM, Caliani C, Palazzi C, Franci N, Perrone S, et al. Inter-observer reliability of two pain scales for newborns. *Early Hum Dev* 2007;83(8):549-552. [doi: [10.1016/j.earlhumdev.2006.10.006](https://doi.org/10.1016/j.earlhumdev.2006.10.006)] [Medline: [17161923](https://pubmed.ncbi.nlm.nih.gov/17161923/)]
12. Pillai Riddell RR, Jasim S, Hamwi L. Out of the mouth of babes: a lot about pain has nothing to do with pain. *Pain* 2022;163(Suppl 1):S117-S125. [doi: [10.1097/j.pain.0000000000002761](https://doi.org/10.1097/j.pain.0000000000002761)] [Medline: [36252235](https://pubmed.ncbi.nlm.nih.gov/36252235/)]
13. Sutton RT, Pincock D, Baumgart DC, Sadowski DC, Fedorak RN, Kroeker KI. An overview of clinical decision support systems: benefits, risks, and strategies for success. *NPJ Digit Med* 2020;3:17 [FREE Full text] [doi: [10.1038/s41746-020-0221-y](https://doi.org/10.1038/s41746-020-0221-y)] [Medline: [32047862](https://pubmed.ncbi.nlm.nih.gov/32047862/)]
14. Cheng D, Liu D, Philpotts LL, Turner DP, Houle TT, Chen L, et al. Current state of science in machine learning methods for automatic infant pain evaluation using facial expression information: study protocol of a systematic review and meta-analysis. *BMJ Open* 2019;9(12):e030482 [FREE Full text] [doi: [10.1136/bmjopen-2019-030482](https://doi.org/10.1136/bmjopen-2019-030482)] [Medline: [31831532](https://pubmed.ncbi.nlm.nih.gov/31831532/)]
15. Matava C, Pankiv E, Ahumada L, Weingarten B, Simpao A. Artificial intelligence, machine learning and the pediatric airway. *Paediatr Anaesth* 2020;30(3):264-268. [doi: [10.1111/pan.13792](https://doi.org/10.1111/pan.13792)] [Medline: [31845543](https://pubmed.ncbi.nlm.nih.gov/31845543/)]
16. Villarroel M, Chaichulee S, Jorge J, Davis S, Green G, Arteta C, et al. Non-contact physiological monitoring of preterm infants in the neonatal intensive care unit. *NPJ Digit Med* 2019;2:128 [FREE Full text] [doi: [10.1038/s41746-019-0199-5](https://doi.org/10.1038/s41746-019-0199-5)] [Medline: [31872068](https://pubmed.ncbi.nlm.nih.gov/31872068/)]
17. Salekin MS, Zamzmi G, Goldgof D, Kasturi R, Ho T, Sun Y. Multimodal spatio-temporal deep learning approach for neonatal postoperative pain assessment. *Comput Biol Med* 2021;129:104150 [FREE Full text] [doi: [10.1016/j.compbiomed.2020.104150](https://doi.org/10.1016/j.compbiomed.2020.104150)] [Medline: [33348218](https://pubmed.ncbi.nlm.nih.gov/33348218/)]
18. Antes AL, Burrous S, Sisk BA, Schuelke MJ, Keune JD, DuBois JM. Exploring perceptions of healthcare technologies enabled by artificial intelligence: an online, scenario-based survey. *BMC Med Inform Decis Mak* 2021;21(1):221 [FREE Full text] [doi: [10.1186/s12911-021-01586-8](https://doi.org/10.1186/s12911-021-01586-8)] [Medline: [34284756](https://pubmed.ncbi.nlm.nih.gov/34284756/)]
19. Asan O, Bayrak AE, Choudhury A. Artificial intelligence and human trust in healthcare: focus on clinicians. *J Med Internet Res* 2020;22(6):e15154 [FREE Full text] [doi: [10.2196/15154](https://doi.org/10.2196/15154)] [Medline: [32558657](https://pubmed.ncbi.nlm.nih.gov/32558657/)]
20. O'Dell B, Stevens K, Tomlinson A, Singh I, Cipriani A. Building trust in artificial intelligence and new technologies in mental health. *Evid Based Ment Health* 2022;25(2):45-46 [FREE Full text] [doi: [10.1136/ebmental-2022-300489](https://doi.org/10.1136/ebmental-2022-300489)] [Medline: [35444002](https://pubmed.ncbi.nlm.nih.gov/35444002/)]
21. Fritsch SJ, Blankenheim A, Wahl A, Hetfeld P, Maassen O, Deffge S, et al. Attitudes and perception of artificial intelligence in healthcare: a cross-sectional survey among patients. *Digit Health* 2022;8:20552076221116772 [FREE Full text] [doi: [10.1177/20552076221116772](https://doi.org/10.1177/20552076221116772)] [Medline: [35983102](https://pubmed.ncbi.nlm.nih.gov/35983102/)]
22. O'Brien BC, Harris IB, Beckman TJ, Reed DA, Cook DA. Standards for Reporting Qualitative Research: a synthesis of recommendations. *Acad Med* 2014 Sep;89(9):1245-1251 [FREE Full text] [doi: [10.1097/ACM.0000000000000388](https://doi.org/10.1097/ACM.0000000000000388)] [Medline: [24979285](https://pubmed.ncbi.nlm.nih.gov/24979285/)]
23. Tie YC, Birks M, Francis K. Grounded theory research: a design framework for novice researchers. *SAGE Open Med* 2019;7:2050312118822927 [FREE Full text] [doi: [10.1177/2050312118822927](https://doi.org/10.1177/2050312118822927)] [Medline: [30637106](https://pubmed.ncbi.nlm.nih.gov/30637106/)]
24. Pillai Riddell RR, Stevens BJ, McKeever P, Gibbins S, Asztalos L, Katz J, et al. Chronic pain in hospitalized infants: health professionals' perspectives. *J Pain* 2009;10(12):1217-1225 [FREE Full text] [doi: [10.1016/j.jpain.2009.04.013](https://doi.org/10.1016/j.jpain.2009.04.013)] [Medline: [19541547](https://pubmed.ncbi.nlm.nih.gov/19541547/)]
25. Huartson K, Hill T, Killam T, Kelly M, Racine N. Physician perspectives on the implementation of a trauma informed care initiative in the maternity care setting. *Int J Child Adolesc Resilience* 2022;9(1):205-215 [FREE Full text] [doi: [10.54488/ijcar.2022.313](https://doi.org/10.54488/ijcar.2022.313)]
26. Guest G, Bunce A, Johnson L. How many interviews are enough? An experiment with data saturation and variability. *Field Methods* 2016;18(1):59-82. [doi: [10.1177/1525822x05279903](https://doi.org/10.1177/1525822x05279903)]
27. Braun V, Clarke V. Using thematic analysis in psychology. *Qual Res Psychol* 2006;3(2):77-101. [doi: [10.1191/1478088706qp063oa](https://doi.org/10.1191/1478088706qp063oa)]

28. McHugh ML. Interrater reliability: the kappa statistic. *Biochem Med* 2012;22(3):276-282 [FREE Full text] [doi: [10.11613/bm.2012.031](https://doi.org/10.11613/bm.2012.031)]
29. Prgomet M, Cardona-Morrell M, Nicholson M, Lake R, Long J, Westbrook J, et al. Vital signs monitoring on general wards: clinical staff perceptions of current practices and the planned introduction of continuous monitoring technology. *Int J Qual Health Care* 2016;28(4):515-521 [FREE Full text] [doi: [10.1093/intqhc/mzw062](https://doi.org/10.1093/intqhc/mzw062)] [Medline: [27317251](https://pubmed.ncbi.nlm.nih.gov/27317251/)]
30. Nash DM, Thorpe C, Brown JB, Kueper JK, Rayner J, Lizotte DJ, et al. Perceptions of artificial intelligence use in primary care: a qualitative study with providers and staff of Ontario Community Health Centres. *J Am Board Fam Med* 2023;36(2):221-228 [FREE Full text] [doi: [10.3122/jabfm.2022.220177R2](https://doi.org/10.3122/jabfm.2022.220177R2)] [Medline: [36948536](https://pubmed.ncbi.nlm.nih.gov/36948536/)]
31. Sisk BA, Antes AL, Burrous S, DuBois JM. Parental attitudes toward artificial intelligence-driven precision medicine technologies in pediatric healthcare. *Children (Basel)* 2020;7(9):145 [FREE Full text] [doi: [10.3390/children7090145](https://doi.org/10.3390/children7090145)] [Medline: [32962204](https://pubmed.ncbi.nlm.nih.gov/32962204/)]
32. Shah N, Arshad A, Mazer MB, Carroll CL, Shein SL, Remy KE. The use of machine learning and artificial intelligence within pediatric critical care. *Pediatr Res* 2023;93(2):405-412 [FREE Full text] [doi: [10.1038/s41390-022-02380-6](https://doi.org/10.1038/s41390-022-02380-6)] [Medline: [36376506](https://pubmed.ncbi.nlm.nih.gov/36376506/)]
33. McCradden MD, Sarker T, Paprica PA. Conditionally positive: a qualitative study of public perceptions about using health data for artificial intelligence research. *BMJ Open* 2020;10(10):e039798 [FREE Full text] [doi: [10.1136/bmjopen-2020-039798](https://doi.org/10.1136/bmjopen-2020-039798)] [Medline: [33115901](https://pubmed.ncbi.nlm.nih.gov/33115901/)]
34. Thai K, Tsiandoulas KH, Stephenson EA, Menna-Dack D, Zlotnik Shaul R, Anderson JA, et al. Perspectives of youths on the ethical use of artificial intelligence in health care research and clinical care. *JAMA Netw Open* 2023;6(5):e2310659 [FREE Full text] [doi: [10.1001/jamanetworkopen.2023.10659](https://doi.org/10.1001/jamanetworkopen.2023.10659)] [Medline: [37126349](https://pubmed.ncbi.nlm.nih.gov/37126349/)]
35. Naik N, Hameed BMZ, Shetty DK, Swain D, Shah M, Paul R, et al. Legal and ethical consideration in artificial intelligence in healthcare: who takes responsibility? *Front Surg* 2022;9:862322 [FREE Full text] [doi: [10.3389/fsurg.2022.862322](https://doi.org/10.3389/fsurg.2022.862322)] [Medline: [35360424](https://pubmed.ncbi.nlm.nih.gov/35360424/)]
36. Lobe B, Morgan DL, Hoffman K. A systematic comparison of in-person and video-based online interviewing. *Int J Qual Methods* 2022;21:1-12 [FREE Full text] [doi: [10.1177/16094069221127068](https://doi.org/10.1177/16094069221127068)]
37. Oftedal A, Busterud K, Irgens LM, Haug K, Rasmussen S. Socio-economic risk factors for preterm birth in Norway 1999-2009. *Scand J Public Health* 2016;44(6):587-592. [doi: [10.1177/1403494816653288](https://doi.org/10.1177/1403494816653288)] [Medline: [27307464](https://pubmed.ncbi.nlm.nih.gov/27307464/)]

Abbreviations

AI: artificial intelligence

HCP: health care professional

MSH: Mount Sinai Hospital

NICU: neonatal intensive care unit

UCLH: University College London Hospital

Edited by K El Emam, B Malin; submitted 02.08.23; peer-reviewed by M Görge, J Haverinen; comments to author 30.09.23; revised version received 24.11.23; accepted 17.12.23; published 09.02.24.

Please cite as:

Racine N, Chow C, Hamwi L, Bucsea O, Cheng C, Du H, Fabrizi L, Jasim S, Johannsson L, Jones L, Laudiano-Dray MP, Meek J, Mistry N, Shah V, Stedman I, Wang X, Riddell RP

Health Care Professionals' and Parents' Perspectives on the Use of AI for Pain Monitoring in the Neonatal Intensive Care Unit: Multisite Qualitative Study

JMIR AI 2024;3:e51535

URL: <https://ai.jmir.org/2024/1/e51535>

doi:[10.2196/51535](https://doi.org/10.2196/51535)

PMID:[38875686](https://pubmed.ncbi.nlm.nih.gov/38875686/)

©Nicole Racine, Cheryl Chow, Lojain Hamwi, Oana Bucsea, Carol Cheng, Hang Du, Lorenzo Fabrizi, Sara Jasim, Lesley Johannsson, Laura Jones, Maria Pureza Laudiano-Dray, Judith Meek, Neelum Mistry, Vibhuti Shah, Ian Stedman, Xiaogang Wang, Rebecca Pillai Riddell. Originally published in JMIR AI (<https://ai.jmir.org>), 09.02.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR AI, is properly cited. The complete bibliographic information, a link to the original publication on <https://www.ai.jmir.org/>, as well as this copyright and license information must be included.

Original Paper

Comparing the Efficacy and Efficiency of Human and Generative AI: Qualitative Thematic Analyses

Maximo R Prescott^{1,2}, MPH; Samantha Yeager¹, PhD; Lillian Ham^{1,2}, MS; Carlos D Rivera Saldana^{1,3}, PhD, MS; Vanessa Serrano^{1,2}, BA; Joey Narez¹, BS; Dafna Paltin^{1,2}, BS; Jorge Delgado¹, BS; David J Moore^{1,4}, PhD; Jessica Montoya^{1,4}, PhD

¹HIV Neurobehavioral Research Program, University of California, San Diego, San Diego, CA, United States

²San Diego State University/University of California San Diego Joint Doctoral Program in Clinical Psychology, San Diego, CA, United States

³Department of Medicine, University of California, San Diego, San Diego, CA, United States

⁴Department of Psychiatry, University of California, San Diego, La Jolla, CA, United States

Corresponding Author:

Maximo R Prescott, MPH

HIV Neurobehavioral Research Program

University of California, San Diego

220 Dickinson Street

San Diego, CA, 92103

United States

Phone: 1 7602713336

Email: mrprescott@health.ucsd.edu

Abstract

Background: Qualitative methods are incredibly beneficial to the dissemination and implementation of new digital health interventions; however, these methods can be time intensive and slow down dissemination when timely knowledge from the data sources is needed in ever-changing health systems. Recent advancements in generative artificial intelligence (GenAI) and their underlying large language models (LLMs) may provide a promising opportunity to expedite the qualitative analysis of textual data, but their efficacy and reliability remain unknown.

Objective: The primary objectives of our study were to evaluate the consistency in themes, reliability of coding, and time needed for inductive and deductive thematic analyses between GenAI (ie, ChatGPT and Bard) and human coders.

Methods: The qualitative data for this study consisted of 40 brief SMS text message reminder prompts used in a digital health intervention for promoting antiretroviral medication adherence among people with HIV who use methamphetamine. Inductive and deductive thematic analyses of these SMS text messages were conducted by 2 independent teams of human coders. An independent human analyst conducted analyses following both approaches using ChatGPT and Bard. The consistency in themes (or the extent to which the themes were the same) and reliability (or agreement in coding of themes) between methods were compared.

Results: The themes generated by GenAI (both ChatGPT and Bard) were consistent with 71% (5/7) of the themes identified by human analysts following inductive thematic analysis. The consistency in themes was lower between humans and GenAI following a deductive thematic analysis procedure (ChatGPT: 6/12, 50%; Bard: 7/12, 58%). The percentage agreement (or intercoder reliability) for these congruent themes between human coders and GenAI ranged from fair to moderate (ChatGPT, inductive: 31/66, 47%; ChatGPT, deductive: 22/59, 37%; Bard, inductive: 20/54, 37%; Bard, deductive: 21/58, 36%). In general, ChatGPT and Bard performed similarly to each other across both types of qualitative analyses in terms of consistency of themes (inductive: 6/6, 100%; deductive: 5/6, 83%) and reliability of coding (inductive: 23/62, 37%; deductive: 22/47, 47%). On average, GenAI required significantly less overall time than human coders when conducting qualitative analysis (20, SD 3.5 min vs 567, SD 106.5 min).

Conclusions: The promising consistency in the themes generated by human coders and GenAI suggests that these technologies hold promise in reducing the resource intensiveness of qualitative thematic analysis; however, the relatively lower reliability in coding between them suggests that hybrid approaches are necessary. Human coders appeared to be better than GenAI at identifying nuanced and interpretative themes. Future studies should consider how these powerful technologies can be best used in collaboration

with human coders to improve the efficiency of qualitative research in hybrid approaches while also mitigating potential ethical risks that they may pose.

(*JMIR AI* 2024;3:e54482) doi:[10.2196/54482](https://doi.org/10.2196/54482)

KEYWORDS

GenAI; generative artificial intelligence; ChatGPT; Bard; qualitative research; thematic analysis; digital health

Introduction

Background

Qualitative methods are pivotal for the development and implementation of digital health interventions. In implementation science, qualitative methods are often used to inform, refine, and improve digital health interventions [1]. Thematic analysis can be applied to qualitative data generated from various methods or sources (eg, key informant interviews and focus groups). This flexible and broad method involves identifying, extracting, and interpreting common themes (ie, codes) within the data that are not subscribed to a particular theory [2,3]. These themes may be identified via inductive (“bottom-up”) or deductive (“top-down”) methods [2]. In the former, themes are data driven, reflecting a rich description of the overall data. In contrast, the latter is driven by existing literature and previously published health behavior models, resulting in a detailed analysis of specific data that fit within a priori coding frames.

Compared to quantitative methods, qualitative methods are often more resource and cost intensive, conflicting with the need for timely feedback in rapidly changing real-world settings (eg, changes in health care policies and patient needs). Such delays in research on evidence-based practices unfortunately minimize their relevance and applicability [4]. An emerging alternative to traditional qualitative methods includes rapid qualitative analyses, which most commonly aim to reduce the time invested in data collection, management, analysis, and interpretation [5,6]. Studies comparing rapid qualitative analyses to traditional methods have shown a good overlap between themes [1,5,7], with additional benefits such as greater data collection and decreased costs [6]. Nonetheless, ongoing challenges to rapid analyses include reduced scientific rigor (ie, trustworthiness) [6,8] and an intensified workload due to a truncated timeline [1,5].

ChatGPT (Open AI) and Google Bard (subsequently rebranded as Gemini) are 2 popular generative artificial intelligence (GenAI)-based systems that provide an interface for humans to collaborate with powerful large language models (LLMs): OpenAI’s GPT-3.5 neural engine [9] and Google’s PaLM 2 [10]; these models are trained to predict and generate humanlike textual responses by leveraging deep learning techniques on massive amounts of pre-existing textual data [11,12]. Recently, LLMs have outperformed previously developed artificial intelligence (AI) systems across different tasks spanning a wide range of disciplines [13]. There is a growing interest in exploring clinical uses for GenAI including new drug design [14] and brain tumor imaging [15]. In the digital health setting, GenAI apps offer individualized information to users on diverse health topics including chronic and infectious diseases or healthy

lifestyle choices [16]. In research, GenAI functions can range from summarizing literature and analyzing data (including coding) to identifying research gaps and drafting papers [17]. Despite these powerful uses, questions remain about the reliability of GenAI as a research tool, given the possibility that GenAI generates incorrect text (eg, “hallucinations”) and distorts scientific facts [17].

In the realm of qualitative research, the interpretation of observed events introduces significant subjectivity. Triangulation [8,18] is a strategy to improve the validity or efficacy of qualitative analysis by integrating information from different sources (eg, human- vs computer-derived codebooks), thereby leveraging the advantages of multiple data analysis methods. For example, Firmin et al [19] found that human-generated thematic codes and software-driven categories were highly correlated for concrete constructs but highlighted unique subjective or abstract constructs. Qualitative analysis by human coders targets meanings and interpretations, whereas LLMs target structural and logical elements of language [11]. To date, there have only been 2 known studies that have recently demonstrated and evaluated the efficacy of applying GenAI for qualitative research compared to human analysts. de Paoli [11] explored whether the LLM underlying ChatGPT could be used to conduct inductive thematic analysis. The results suggested that at least some of the themes previously identified by human analysts in the contexts of education and psychology were able to be reproduced by GenAI and warranted further exploration and methodological considerations [11]. Alternatively, Hamilton et al [20] leveraged ChatGPT to conduct a phenomenological qualitative analysis of significant statements from the interview transcripts nested within a guaranteed income program evaluation and compared its identified themes to those generated by human analysts. They similarly found promising similarities in identified themes as well as discrepancies such as limited contextual understanding from ChatGPT.

Although substantial debate remains as to how to best evaluate the methodological rigor or trustworthiness of qualitative research, the accuracy in which findings reflect the data (ie, efficacy) and the reliability within analytic procedures are prominent considerations [21]. The extent to which GenAI can produce rigorous and trustworthy qualitative research while reducing the time and resource burden of current qualitative methods remains open to exploration, particularly within the context of health-related research.

Objectives

The primary objectives of our study were to assess the consistency and reliability of thematic analysis conducted by ChatGPT, Bard, and human coders following both inductive and deductive approaches. In this paper, we have described and

compared the methods that we used among humans and GenAI to contribute to the growing body of literature on the wide-ranging applications of GenAI for qualitative analysis in digital health research. Specifically, we aimed to compare both the consistency in identifying broad themes essential to qualitative research and the reliability in coding between methods (ie, humans, ChatGPT, and Bard). Furthermore, we additionally examined the difference in human resources required (ie, time spent on the analysis) between the methods for both approaches.

Methods

Qualitative Data

The qualitative data for this study consisted of 40 short (<160 characters; 5-14 words in length) SMS text message prompts used in a previous study evaluating an SMS text messaging intervention (individualized texting for adherence building; iTAB) to promote antiretroviral medication adherence among people with HIV who use methamphetamine [22]. The iTAB messages draw from various health behavior models including the health belief model [23], theory of planned behavior [24], social cognitive theory [25], and attitude–social influence–efficacy model [26]. During the development of iTAB, sample messages were tested among people with HIV who provided feedback; participant feedback was subsequently used to adapt the SMS text messages. These SMS text messages served as the foundation for the final, streamlined version [27]. For this study, the 40 short SMS text messages were the qualitative data being analyzed. We used SMS text message prompts as opposed to participant-generated qualitative responses in our analyses, as GenAI services record all data entered to further train LLMs.

Ethical Considerations

Presently, the use of participant-generated SMS text messages would violate the protection of confidentiality agreements as

the consent forms approved by our institutional review board did not specify that participant-generated qualitative data would be uploaded to third-party vendors. We believe that using the 40 short SMS text message prompts provides a proxy for qualitative data to model how GenAI and LLMs compare in detecting shared themes among the SMS text messages compared to human-conducted thematic analysis. This was an institutional review board–exempt study as there were no data from human participants involved.

GenAI Services

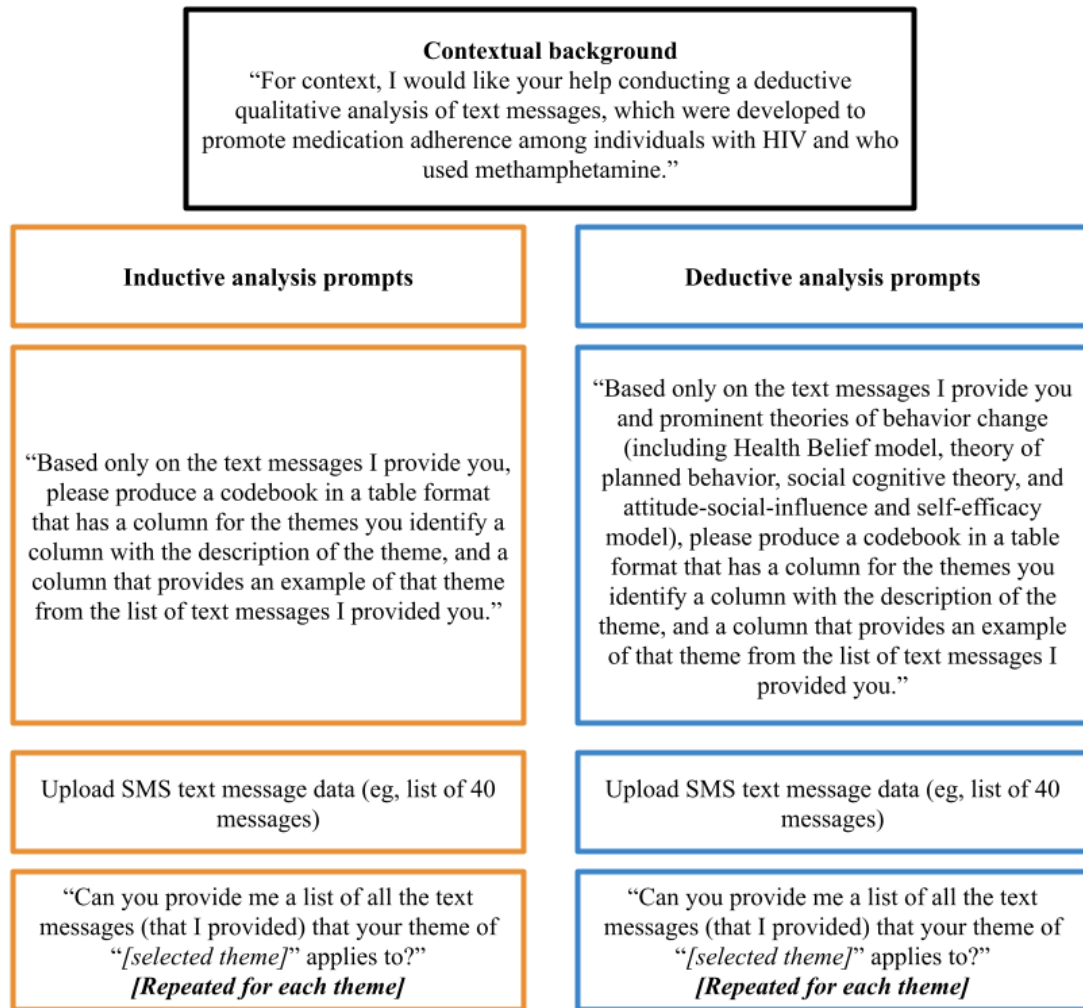
Overview

We used 2 commercially available GenAI services: ChatGPT-3.5 (OpenAI) and Bard (Google). ChatGPT-3.5 leverages OpenAI's proprietary LLM (GPT-3.5), which was trained using reinforcement learning from human feedback [28], a method that provides rewards to reinforce learning. Bard is powered by Google's proprietary LLM (PaLM 2) [10], a transformer-based model that enables it to conduct advanced reasoning tasks including classification and language generation. Both LLMs are currently free and open to the public.

GenAI: Inductive Thematic Analysis Procedures

ChatGPT and Bard were given identical prompts to conduct inductive thematic analysis. Before providing the SMS text message prompts, both GenAI services were prompted with contextual information on the study and a description of the procedures, which the independent human analyst asked GenAI to perform (Figure 1). Following this contextual information and instructions, the SMS text messages were copied into the GenAI interface all at once (ie, as a list of 40 messages), and the model exported the requested 3-column codebook table. We then provided additional instructions to have the GenAI label the SMS text messages based on shared themes (Figure 1).

Figure 1. Generative artificial intelligence thematic analysis instruction prompts.



GenAI: Deductive Thematic Analysis Procedures

ChatGPT and Bard were given identical prompts to conduct deductive thematic analysis. Both GenAI services were prompted with the same background contextual information and instructions as the inductive thematic analysis prompts. All SMS text messages for analyses were similarly copied all at once into the GenAI. However, the deductive approach additionally requested that SMS text messages be categorized using constructs from relevant theories of behavior change (such as medication adherence) including the health belief model [23], theory of planned behavior [24], social cognitive theory [25], and attitude–social influence–efficacy model) [26] (Figure 1).

Training of Human Coders and Analyst

In terms of training, all the 4 human coders responsible for the qualitative analysis had been previously trained by the senior author on the proper conduct of qualitative analysis on prior studies, as well as had attended formal external webinars on qualitative coding and analysis. Each thematic analysis was conducted by 2 research team members consisting of a clinical psychology doctoral student and a research assistant with a bachelor’s degree. These research team members were not involved in the development or evaluation of the SMS text messaging intervention. The separate inductive and deductive human teams were thus intentionally balanced in their

experience and expertise with qualitative analysis and were instructed not to discuss or collaborate on their analyses to maintain the independence of each analytic approach.

Similar to human coders, the human analyst responsible for developing the GenAI prompts used in this study had previous training and experience with qualitative analysis. The human analyst also had several years of experience working in technology development, applying emerging technologies to digital health, and incorporating general guidelines on prompt engineering in the context of health care [29]. Specifically, the analyst incorporated guidelines such as being specific, providing the setting and context, identifying the overall goal first, and requesting examples to inform their prompts.

Procedures for Human Inductive Thematic Analysis

For inductive thematic analysis, both members were given contextual background on the SMS text messaging study and were instructed to independently develop their own codebook. Once each member developed an initial codebook, they were instructed to come to a consensus on a final codebook. Following agreement on a final codebook, each team member applied the codebook to group the SMS text messages and began to search for themes. Finally, both team members compared their application of the final codebook, resolved disagreements in coding, and ultimately came to a consensus on the broader

themes derived by summarizing and collapsing codes. This process was consistent with the basic steps of a thematic analysis [2].

Procedures for Human Deductive Thematic Analysis

For deductive thematic analysis, both research members were given the same contextual study information as the inductive thematic analysis team. The deductive analysis team was then given a list of a priori codes based on the theories of behavior change including health belief model [23], theory of planned behavior [24], social cognitive theory [25], and attitude–social influence–efficacy model [26]. Next, the deductive analysis team was instructed to independently develop a codebook considering the key constructs of the theories of behavior change and was suggested a priori codes. Once both team members independently developed a codebook, they were instructed to compare codebooks and reached a consensus on a final codebook. Following agreement on a final codebook, each team member applied the codebook to group the text messages and began searching for themes. Both team members then compared their application of the final codebook, resolved any discrepancies in coding, and came to a consensus on broader themes by collapsing and summarizing their codes. This process was consistent with the basic steps of a thematic analysis [2].

Consistency and Intercoder Reliability

Consistency was defined as the extent to which the thematic findings were the same across the 2 analytic methods, as has previously been used when comparing qualitative methods [1]. We operationalized this as the percentage of themes that was shared between methods (eg, 100% consistency would suggest that the themes between the methods were identical). For example, if method A (reference method) were to identify 10 total themes and method B identified 5 of those themes, then the theme consistency would be 50% (5/10).

Intercoder reliability (ICR) was operationalized as the number of agreements in coding divided by the sum of agreements and disagreements in coding [30]; thus, a higher score equates to a greater agreement between coders. After reaching a final consensus on the codebook, an ICR was calculated. Human coders then met to discuss disagreements in the coding of the data. To compare the reliability of coding between humans and GenAI for themes that were shared by both methods, ICRs were calculated to determine the reliability between (1) human coders and Bard, (2) human coders and ChatGPT, and (3) ChatGPT and Bard. ICRs were averaged across all common themes to compute an overall ICR percentage. To qualitatively describe the extent of agreement between human and GenAI teams, the following cutoffs were used to interpret the ICRs: slight (0%-20%), fair (21%-40%), moderate (41%-60%), substantial (61%-80%), and almost perfect (81%-100%) [31]. The consistency and ICR between methods were descriptively reported and compared.

Total Time Spent on the Analyses

After all thematic analyses had been completed, each human coder was asked to retrospectively estimate the amount of total time spent on their qualitative analyses. For the human coders, this total time included the sum of the time spent by each

individual coder on their initial coding, codebook development, application of codebook, and meetings to reach a consensus on disagreements in coding and themes. The total time spent on thematic analyses using GenAI was the sum of the time taken to input the prompts, waiting for responses to generate, and compiling those responses in a spreadsheet table. The differences in the total time spent on analysis between methods were also descriptively reported and compared.

Results

Consistency of Inductive Thematic Analyses

In the inductive arm of our study, 7 themes were identified by the human coders following an inductive thematic analysis of the iTAB SMS text messages. These themes included “time,” “adherence,” “religious,” “community care,” “health reminder,” “warning,” and “encouragement.” Of these 7 themes identified by human coders, 5 (71%) were also consistent with the themes derived by both ChatGPT and Bard. [Multimedia Appendix 1](#) presents a complete mapping of the inductive thematic analysis codebooks (including theme, description, and example text messages) generated by human coders, ChatGPT, and Bard.

ChatGPT’s inductive thematic analysis of the same SMS text messages identified 10 themes, which included “spirituality/higher power,” “supportive community,” “love and support from others,” “health benefits,” “reminder,” “resistance and risk to others,” “consequences of nonadherence,” “positive reinforcement,” “fun and enjoyment,” and “accountability.” Of these 10 themes, almost all (n=9, 90%) were consistent with the themes identified by our human coders. Bard’s inductive thematic analysis identified 6 themes from the SMS text messages, which included “religious/spiritual beliefs,” “social support,” “importance of taking medication,” “consequences of not taking medication,” “enjoyment,” and “personal responsibility.” Of these 6 themes, the majority (n=5, 83%) were consistent with the themes identified by human coders, and there was perfect consistency (6/6, 100%) between the themes identified by ChatGPT and Bard.

The 1 theme that ChatGPT identified that human coders did not was “accountability,” which was defined as “Messages emphasizing personal responsibility for adherence.” Bard similarly identified this theme as “personal responsibility,” which was defined as “The messages emphasize that it is important for people to take care of themselves and take their medication on their own. They also suggest that people should be proud of themselves for being adherent to their medication regimen.” The example SMS text messages provided by ChatGPT and Bard that are representative of the “accountability” and “personal responsibility” theme were as follows:

Stop screwing around and take ur [medication] now.
[ChatGPT]

It's imp't to take care of urself. Pls take ur [medication] [Bard]

ChatGPT and Bard did not reproduce 2 (29%) of the 7 human coders’ themes, “time” and “adherence,” which were defined rather literally as including the words “time” or “adherence” in the message. There were 4 instances where 2 themes derived

by ChatGPT ultimately mapped onto a single broader theme identified by the human coders' thematic analysis. For example, the human coders identified "community care," which they defined as "message focused on the importance of the individual in relation to others, both being cared for by others and being accountable to others." By examining both themes and their descriptions, we observed that there were 2 themes identified by ChatGPT that mapped onto "community care" as defined by our human coders:

1. *Supportive community*: messages highlighting the care and support from others
2. *Love and support from others*: messages emphasizing the impact on loved ones

Reliability of Inductive Thematic Coding

The overall ICR of all inductive themes shared by human coding and ChatGPT was moderate (31/66, 47%). The overall ICR was lower between human coders and Bard at 37% (20/54), which is indicative of fair agreement. There was similarly fair agreement in coding between ChatGPT and Bard at 37% (23/62). There was notable variation in the ICR between coding arms when examined by theme, which ranged from 8% (2/26; slight agreement for "encouragement" between ChatGPT and Bard) to 80% (4/5; substantial agreement for "religious" between human coders and ChatGPT, human coders and Bard, and ChatGPT and Bard). [Table 1](#) lists ICR between human coders, ChatGPT, and Bard for inductive thematic coding, both overall and by theme.

Table 1. Inductive thematic analysis intercoder reliability (ICR) between human coders, ChatGPT, and Bard by theme and overall.

Themes	Human coders and Bard		Human coders and ChatGPT		Bard and ChatGPT	
	Agreement, n/N (%; ICR)	Disagreement, n/N (%)	Agreement, n/N (%; ICR)	Disagreement, n/N (%)	Agreement, n/N (%; ICR)	Disagreement, n/N (%)
Encouragement	2/17 (12)	15/17 (88)	12/29 (41)	17/29 (59)	2/26 (8)	24/26 (92)
Health reminder	6/17 (35)	11/17 (65)	7/18 (39)	11/18 (61)	9/15 (60)	6/15 (40)
Religious	4/5 (80)	1/5 (20)	4/5 (80)	1/5 (20)	3/5 (60)	2/5 (40)
Community or cared by others	5/6 (83)	1/6 (17)	4/6 (67)	2/6 (33)	4/5 (80)	1/5 (20)
Warning	3/9 (33)	6/9 (67)	4/8 (50)	4/8 (50)	4/6 (67)	2/6 (33)
Personal responsibility (Bard and ChatGPT only)	— ^a	—	—	—	1/5 (20)	4/5 (80)
Overall (across all themes)	20/54 (37)	34/54 (63)	31/66 (47)	35/66 (53)	23/62 (37)	39/62 (63)

^aNot applicable.

Consistency of Deductive Thematic Analysis

A total of 12 themes were identified by the human coders following a deductive thematic analysis of the same text messages, which included "positive tone," "stern/serious tone," "sense of urgency/priority," "balancing health with 'fun'," "self-care," "expectations and attitudes," "perceived negative outcomes" "perceived benefits," "norms," "social influence," "self-efficacy," and "spirituality/religion as motivation." Of these 12 themes identified by human coders, 6 (50%) were also consistent with the themes derived by ChatGPT, and 7 (58%) were consistent with those found by Bard. [Multimedia Appendix 2](#) lists the complete mapping of the deductive thematic analysis codebooks (including theme, description, and example SMS text messages) generated by human coders, ChatGPT, and Bard.

ChatGPT's deductive thematic analysis identified a total of 9 themes, which included "consequences," "health benefits," "motivation," "social influence," "care and support," "self-efficacy," "religious beliefs," "responsibility," and "reminders." Of these 9 themes, the majority (7/9, 78%) were consistent with the themes identified by our human coders. Bard's inductive thematic analysis identified 6 themes from the SMS text messages, which included "importance of adherence," "negative consequences of nonadherence," "benefits of

adherence," "social support," "self-efficacy," and "religious/spiritual." Of these 6 themes identified by Bard, there was perfect consistency (6/6, 100%) with the themes identified by human coders, and there was strong consistency (5/6, 83%) with the themes identified by ChatGPT.

ChatGPT and Bard did not reproduce 6 (50%) and 5 (42%), respectively, of the human coder's 12 deductive themes. Neither ChatGPT nor Bard identified the human coder's themes of "positive tone," "stern/serious tone," "balancing health with 'fun'," "self-care," or "expectations and attitudes." In addition, ChatGPT did not identify the human coder's theme of "sense of urgency/priority," which was defined as "includes messages instructing a person to place their health, or desired health behaviors, over other competing priorities." On the basis of this description, the Bard theme of "importance of adherence" was mapped onto this theme, and both shared the example message of "Stop everything and take ur meds!"

There were 2 themes identified uniquely by ChatGPT (ie, neither Bard nor human coders identified these themes), which were "responsibility" and "reminders" and were defined as follows:

- **Responsibility**: encouraging a sense of responsibility for one's health and well-being through adherence.

- Example text: *It's imp't to take care of urself. Pls take ur [medication]*
- Reminders: providing reminders or cues to prompt medication adherence.
 - Example text: *Ready, set, get healthy! It's med time. Time for ur [medication]*

There was 1 case where Bard identified a single theme (“social support”) that human coders and ChatGPT had separated into 2 separate themes (“norms” or “social influence” and “social influence” or “care and support,” respectively). Furthermore, there was an instance where 2 themes derived by ChatGPT ultimately mapped onto a broader theme identified by the thematic analysis performed by the human coders and Bard. For example, the human coders identified “perceived benefits,” which they defined as “perception of the effectiveness of an action to reduce the threat of illness or disease, including factors related to ease of use.” By examining both themes and their descriptions, there were 2 ChatGPT themes that mapped onto this theme of “perceived benefits”:

1. *Health benefits*: highlighting the positive impact of medication adherence on health and well-being
2. *Motivation*: encouraging individuals to take their medication by emphasizing the benefits of doing so

Reliability of Deductive Thematic Coding

The overall ICR of deductive themes shared between human coders and ChatGPT was fair at 37% (22/59), which was similar to Bard at 36% (21/58). There was moderate agreement in coding between the codebooks generated by ChatGPT and Bard, as reflected by an overall ICR of 47% (22/47). We also examined code-specific ICR in addition to the overall ICR, which varied substantially across themes. For example, there was perfect (4/4, 100%) agreement in coding between human coders and ChatGPT, human coders and Bard, and ChatGPT and Bard within the theme of “perceived negative outcomes,” but only slight to fair agreement for the theme of “perceived benefits” (6/29, 21%; 5/13, 38%; and 9/28, 32%, respectively). [Table 2](#) presents the ICR between human coders, ChatGPT, and Bard deductive thematic coding, both overall and by theme.

Table 2. Deductive thematic analysis intercoder reliability (ICR) between human coders, ChatGPT, and Bard by theme and overall.

Themes	Human coders and Bard		Human coders and ChatGPT		Bard and ChatGPT	
	Agreement, n/N (%; ICR)	Disagreement, n/N (%)	Agreement, n/N (%; ICR)	Disagreement, n/N (%)	Agreement, n/N (%; ICR)	Disagreement, n/N (%)
Perceived benefits	5/13 (38)	8/13 (62)	6/29 (21)	23/29 (79)	9/28 (32)	19/28 (68)
Perceived negative outcomes	4/4 (100)	0/4 (0)	4/4 (100)	0/4 (0)	4/4 (100)	0/4 (0)
Social Support (Norms and Social influence)	4/14 (29)	10/14 (71)	5/14 (36)	9/14 (64)	4/5 (80)	1/5 (20)
Self-efficacy	2/5 (40)	3/5 (60)	2/7 (29)	5/7 (71)	1/5 (20)	4/5 (80)
Spirituality or religion as motivation	4/5 (80)	1/5 (20)	5/5 (100)	0/5 (0)	4/5 (80)	1/5 (20)
Sense of urgency or priority (Bard only)	2/17 (12)	15/17 (88)	— ^a	—	—	—
Overall (across all themes)	21/58 (36)	37 (64)	22/59 (37)	37/59 (63)	22/47 (47)	25/47 (53)

^aNot applicable.

Total Time Spent on Qualitative Analyses

The human coding teams reported 492 (inductive) and 705 (deductive) total minutes to complete their thematic analyses of the SMS text messages. This total time includes the sum of the time spent by each individual coder on their initial coding, codebook development, application of codebook, and reaching consensus on disagreements in coding. The total time to complete the inductive and deductive thematic analyses with ChatGPT was 15 minutes (97% less time than the human approach) and 25 minutes (97% less), respectively, whereas both analyses took a total of 20 minutes with Bard (96% and 97% less time, respectively). The total time spent on thematic analyses using GenAI was the sum of the time taken to input

the prompts, wait for responses to generate, and document those responses in a spreadsheet table.

Discussion

Principal Findings

This study evaluated the consistency and ICR in themes between human coders and GenAI models conducting both inductive and deductive thematic analyses of short SMS text message prompts that were used in a previous intervention to promote medication adherence. There was evidence of consistency in the themes identified by ChatGPT and Bard compared to human coders' inductive thematic analysis (both 5/7, 71%), but the consistency was notably lower for deductive thematic analysis

(6/12, 50% and 7/12, 58%, respectively). The overall ICR (percent agreement in coding) of themes shared between human coders and GenAI models (inductive: 31/66, 47% and 20/54, 37%; deductive: 22/59, 37% and 21/58, 36%, respectively) was fair to moderate [31]. In addition, GenAI models were significantly less resource-intensive, as they took an average of 97% less time (20 vs 567 min) for qualitative analysis compared to human coders. ChatGPT and Bard performed similarly to each other across both types of thematic analysis.

This study is the first of our knowledge to compare the GenAI- and human-generated themes from textual data following both inductive and deductive qualitative thematic analysis procedures using health-related data. We also evaluated and compared both ChatGPT and Bard, whereas prior studies of GenAI have only examined ChatGPT. Our findings demonstrate that GenAI may provide a promising opportunity to facilitate quicker and more resource-efficient qualitative analysis of textual data; however, such technologies should be used to assist human coders in order to further improve the efficacy and reliability of findings.

Comparison With Prior Work

Although we did not find perfect consistency in AI- and human-generated themes, there were notable similarities in the themes derived by both methods. Hamilton et al [20] similarly compared emergent ChatGPT- and human-generated themes from a qualitative analysis of interview data from a guaranteed income program evaluation, in which they also found an overlap between the 2 methods. They found that approximately 50% of human-generated themes were consistent with those identified by ChatGPT and that 80% of themes identified by ChatGPT were identified by human coders. Furthermore, de Paoli [11] emulated inductive thematic analysis of a previously analyzed semistructured interview data set using the underlying natural language processing (NLP) model of ChatGPT (GPT 3.5-Turbo) and found that a majority of the original themes (9/13, 69%) were identified. The consistency between our GenAI- (ChatGPT and Bard) and human-generated thematic analyses (50%-71%) was notably similar to that observed in these studies (50%-80%) [11,20]. The results of this study and the study by Hamilton et al [20] both found that both GenAI- and human-generated themes were promisingly similar, but both methods also identified distinct themes. Although de Paoli [11] and Hamilton et al [20] had previously demonstrated and evaluated the potential efficacy of GenAI for qualitative research, our findings further suggest that GenAI may also have promising applications for qualitative research in the context of health research.

In this study, the data set consisted of SMS text messages to promote HIV medication adherence for individuals who use methamphetamine and included nuanced references that are unique to this population such as references to substance use (eg, “fun” and “partying”) and specific slang for methamphetamine (“Tina”). The deductive human coding team identified the theme of “balancing health with ‘fun’” based on these messages and recognized the nuance of the use word “fun” in this context as a subtle reference to substance use (codebook description: “Messages contain content reminding a person to prioritize health, even engaging in ‘fun’ or ‘partying behaviors,’ which may include risky behaviors”), whereas ChatGPT and

Bard did not. For these messages, both AI methods tended to take a literal meaning and labeled these messages as representing themes of “fun and enjoyment” or “enjoyment.” However, it is also important to recognize that our inductive coding team similarly did not appear to recognize these subtle references to substance use behaviors. In terms of other notable discrepancies in themes, the human-generated deductive themes included “positive tone” and “stern/serious tone,” which neither ChatGPT nor Bard produced. These themes appear to be consistent with sentiment analysis (or recognizing the sentiment or emotion expressed in text), which is surprising given that recent research has found ChatGPT to be quite promising in sentiment classification of textual data (>92% accuracy) and superior to other NLP methods [32].

One possible explanation for the difference in the results between methods is that GenAI methods appear to be relatively limited in their ability to understand the contextual or subtle meaning of textual data, as they rely primarily on probabilistic pattern recognition to generate responses. The training sets used to train the NLP models underlying ChatGPT and Bard (ie, largely internet content) presumably did not contain a substantial amount of textual data specific to substance use, and so the ability to recognize subtle nuances and references within this context is more limited. The implications of these findings suggest that GenAI shows promise in qualitative thematic analysis but may ultimately prove less valid for less mainstream research topics that may relatively use more nuanced language (eg, illicit substance use), which further highlights the importance of continued inclusion of human coders in the qualitative research process. This possible limitation regarding the more explicit interpretation of GenAI that we observed appears to extend to not only the output it produces but also its use of input (prompts). Whereas our human coders’ understanding of qualitative thematic analysis included the possibility of themes emerging related to the sentiment of the text messages, it appears that ChatGPT and Bard did not. This finding highlights the relative importance of prompt engineering (ie, research into how best to instruct such technologies) and further stresses the importance of maintaining human coders in the qualitative research process when leveraging GenAI.

Currently, there is considerable heterogeneity in the prompts being used to conduct qualitative research with GenAI to date, such that some examples have had the models show their work and used step-by-step prompts following a typical 6-step process [11], and others have been more global in their approach [20] similar to ours. While general guidance exists on how best to prompt GenAI in the context of health care [29], we believe that this represents a critical future direction of research for this field of work. As more studies and case examples are published, a systematic review would be critical for developing best practices and standards for prompt engineering in the specific context of qualitative analysis (eg, are step-by-step prompts better received than more global ones such as the ones used in our study?).

Given that GenAI and human coders were operating off independently derived codebooks and themes, the degree of disagreement in ICR is not surprising. After the initial application of their codebooks and before meeting and coming

to a consensus on disagreements, our inductive and deductive human coding teams had ICRs of 83% (57/69) and 31% (45/144), respectively. Therefore, the ICRs found between our human coders before discussing disagreements was in fact quite similar to the ICRs we observed between human and GenAI methods for shared themes. Furthermore, Xiao et al [33] similarly examined the degree of ICR between a pretrained LLM (GPT-3) and deductive coding conducted by expert human coders, in which they also found fair to substantial agreement between methods [33]. Regarding the difference in ICRs between the inductive and deductive analyses, current literature suggests that this most likely is a reflection of the method and the associated number of codes. In our study, the human deductive analysis team identified more themes (and codes) than did the inductive analysis team (12 vs 7, respectively), which could be suspected as they had been provided with a priori codes from several theories of behavior change. Previous research has found that a greater number of codes reduces the ICR [34,35], which is believed to reflect having to be familiar with a relatively longer coding scheme and thus being more cognitively taxing [36]. There is also substantial debate over the utility of ICR in qualitative methods, as some argue that the inherent subjectivity of qualitative research and the resulting researcher's reflexivity and personal engagement are necessary for understanding the diversity of perspectives on a given topic rather than treating it as noise to be minimized [36,37]. The arguments in support of using ICR are that it helps ensure that themes and information being derived from qualitative data are consistent and meaningful [38]. Therefore, GenAI-generated qualitative analyses may be useful as tools for providing an additional perspective of the data to complement those found by human coders and enabling triangulation and recognition of potential biases.

A significant barrier to qualitative analysis is the considerable time and resources involved, which can be particularly salient when rapid research findings are urgently needed to improve the dissemination and implementation of evidence-based health interventions [4,39]. Previous innovations in qualitative methods, such as rapid qualitative analysis, have shown promise in helping maintain the rigor of the analysis while being quicker and more cost-efficient than traditional methods [1,7]. However, such methodologies still require substantial human resources, and the resource efficiency of qualitative methodologies may be further improved when augmented with new technologies. Several examples of hybrid NLP-qualitative methods, whereby human coders and NLP or GenAI technologies collaborate during analysis, have been proposed or demonstrated previously [40-43]. Skeen et al [41] have provided one such example of a hybrid approach in their proof-of-concept study that applied NLP to condense a large data set of unstructured textual data before subsequent human-generated thematic analysis in order to more rapidly produce design insights for improving a digital HIV intervention.

However, most of these studies using hybrid methods have only demonstrated proof of concept and lacked comparisons with gold-standard qualitative analysis conducted by human analysts. In one previous study comparing qualitative analysis with human coders, NLP-only, and NLP-hybrid methods, the authors found

similar thematic findings across methods and that NLP and hybrid methods required notably less time and resources [43]. Whereas the technical skills (eg, coding) required to implement NLP methods previously posed a significant barrier to the wider adoption of such methodologies among qualitative researchers, commercially available GenAI services, such as ChatGPT and Bard, provide a promising opportunity for further exploration of hybrid NLP-qualitative methods. An example of a hybrid approach incorporating GenAI might be for it to complete the often time-intensive initial coding of textual data, which could subsequently be reviewed and summarized by human analysts to produce the themes that are often more interpretative and abstract in nature. Alternatively, a single human coder might conduct a complete thematic analysis and then collaborate with GenAI as if they were another human coder to reflect on discrepancies and convergence between their coding and identification of broader themes (ie, replacing the need for a second human coder or analyst). The unknown feasibility, efficacy, and efficiency of such hybrid approaches leveraging GenAI warrant future exploration and study.

Limitations

There are several important limitations to consider when interpreting the findings of our study and more broadly the application of GenAI to qualitative analysis. First and foremost, there are numerous current ethical and privacy issues to applying GenAI to human participant research. These issues are currently being debated as these technologies emerge and include the potential for perpetuating bias and inequality, fact fabrication, plagiarism, and potential breaches of data privacy or ownership [44-48]. Using GenAI in the research process also poses potential challenges to obtaining informed consent from participants, especially when working with at-risk populations such as those living with HIV or those who use substances. Obtaining informed consent is fundamental to the ethical conduct of research and involves disclosing to potential participants how their data could be used and the risks associated with research participation, both of which may be difficult to do in the context of using GenAI services due to their lack of transparency, explainability (eg, black box), and the potential risk of reidentification [49]. Future researchers should continue to prudently investigate and monitor both the potential benefits and risks associated with groundbreaking technologies such as GenAI services, especially when incorporating them into the scientific process and their use with data from vulnerable populations.

In addition, it is important to note that our data set consisted of only relatively brief SMS text message prompts that could easily be provided to ChatGPT and Bard. We do not know how well the consistency and reliability of themes derived by GenAI would compare to human coders for longer textual data sets that are common in qualitative research (ie, unstructured or structured interview, focus group transcripts, etc). Our data set also notably did not consist of natural, participant-generated language (eg, transcribed spoken language in interviews), so our findings may not generalize to these more likely data sets for qualitative analysis. However, recent studies have conducted qualitative analyses using GenAI services or their underlying LLMs with such data sets (eg, unstructured qualitative interview

transcripts and significant statements from transcripts) and have shown promisingly similar results to ours [11,20].

Relatedly, we observed challenges with GenAI being able to recognize interpretative themes and consider the nuanced meaning of some topics (specifically, substance use), which may also suggest that our findings may not generalize to all research content areas. Although the flexibility of thematic analysis allows and expects to some degree that initial codes go on to form main themes [2,50], our study was relatively limited in the extent to which we could determine whether GenAI correctly identifies more complex themes due to our small data set of brief text messages. Given the limited research that exists examining the consistency and reliability of applying GenAI to qualitative research and the novelty of the field, future studies should consider further exploring how well these methods generalize to other types and content of data.

Conclusions

Our findings suggest that GenAI may have promising applications for qualitative thematic analysis (including reducing the time and resources required), but hybrid approaches that allow for collaboration between human coders and GenAI technologies are likely necessary to further improve the consistency and reliability of such methods. Improvements in efficiency may be particularly important to further facilitating the adoption of qualitative methods for studying and improving digital health interventions within often complex and rapidly changing real-world settings. As GenAI models are expected to continually improve as they learn, future studies should further explore how humans can best collaborate with these powerful tools given their potential for enabling more rapid research while also remaining vigilant of the potential risks they may pose. Research into the ethical challenges posed by GenAI in the context of human participant research is also urgently needed.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Inductive thematic analysis codebooks generated by human coders, ChatGPT, and Bard.

[\[DOCX File, 26 KB - ai_v3i1e54482_app1.docx\]](#)

Multimedia Appendix 2

Deductive thematic analysis codebooks generated by human coders, ChatGPT, and Bard.

[\[DOCX File, 26 KB - ai_v3i1e54482_app2.docx\]](#)

References

1. Gale RC, Wu J, Erhardt T, Bounthavong M, Reardon CM, Damschroder LJ, et al. Comparison of rapid vs in-depth qualitative analytic methods from a process evaluation of academic detailing in the Veterans Health Administration. *Implement Sci* 2019 Feb 01;14(1):11 [FREE Full text] [doi: [10.1186/s13012-019-0853-y](https://doi.org/10.1186/s13012-019-0853-y)] [Medline: [30709368](https://pubmed.ncbi.nlm.nih.gov/30709368/)]
2. Braun V, Clarke V. Using thematic analysis in psychology. *Qual Res Psychol* 2006 Jan;3(2):77-101. [doi: [10.1191/1478088706qp063oa](https://doi.org/10.1191/1478088706qp063oa)]
3. Lester J, Cho Y, Lochmiller C. Learning to do qualitative data analysis: a starting point. *Hum Resour Dev Rev* 2020 Feb 09;19(1):94-106. [doi: [10.1177/1534484320903890](https://doi.org/10.1177/1534484320903890)]
4. Glasgow RE, Chambers D. Developing robust, sustainable, implementation systems using rigorous, rapid and relevant science. *Clin Transl Sci* 2012 Feb;5(1):48-55 [FREE Full text] [doi: [10.1111/j.1752-8062.2011.00383.x](https://doi.org/10.1111/j.1752-8062.2011.00383.x)] [Medline: [22376257](https://pubmed.ncbi.nlm.nih.gov/22376257/)]
5. Taylor B, Henshall C, Kenyon S, Litchfield I, Greenfield S. Can rapid approaches to qualitative analysis deliver timely, valid findings to clinical leaders? a mixed methods study comparing rapid and thematic analysis. *BMJ Open* 2018 Oct 08;8(10):e019993 [FREE Full text] [doi: [10.1136/bmjopen-2017-019993](https://doi.org/10.1136/bmjopen-2017-019993)] [Medline: [30297341](https://pubmed.ncbi.nlm.nih.gov/30297341/)]
6. Vindrola-Padros C, Johnson GA. Rapid techniques in qualitative research: a critical review of the literature. *Qual Health Res* 2020 Aug;30(10):1596-1604. [doi: [10.1177/1049732320921835](https://doi.org/10.1177/1049732320921835)] [Medline: [32667277](https://pubmed.ncbi.nlm.nih.gov/32667277/)]
7. Nevedal AL, Reardon CM, Opra Widerquist MA, Jackson GL, Cutrona SL, White BS, et al. Rapid versus traditional qualitative analysis using the Consolidated Framework for Implementation Research (CFIR). *Implement Sci* 2021 Jul 02;16(1):67 [FREE Full text] [doi: [10.1186/s13012-021-01111-5](https://doi.org/10.1186/s13012-021-01111-5)] [Medline: [34215286](https://pubmed.ncbi.nlm.nih.gov/34215286/)]
8. McNall M, Foster-Fishman PG. Methods of rapid evaluation, assessment, and appraisal. *Am J Eval* 2016 Jun 30;28(2):151-168. [doi: [10.1177/1098214007300895](https://doi.org/10.1177/1098214007300895)]
9. Researcher Access Program application. OpenAI. URL: <https://openai.com/form/researcher-access-program/> [accessed 2023-09-22]
10. Anil R, Dai AM, Firat O, Johnson M, Lepikhin D, Passos A, et al. PaLM 2 technical report. arXiv Preprint posted online on May 17, 2023 [FREE Full text] [doi: [10.1038/protex.2013.081](https://doi.org/10.1038/protex.2013.081)]

11. de Paoli S. Can large language models emulate an inductive thematic analysis of semi-structured interviews? an exploration and provocation on the limits of the approach and the model. arXiv Preprint posted online on May 22, 2023 [[FREE Full text](#)] [doi: [10.1177/08944393231220483](https://doi.org/10.1177/08944393231220483)]
12. Cooper G. Examining science education in ChatGPT: an exploratory study of generative artificial intelligence. *J Sci Educ Technol* 2023 Mar 22;32(3):444-452. [doi: [10.1007/s10956-023-10039-y](https://doi.org/10.1007/s10956-023-10039-y)]
13. Biever C. ChatGPT broke the Turing test - the race is on for new ways to assess AI. *Nature* 2023 Jul;619(7971):686-689. [doi: [10.1038/d41586-023-02361-7](https://doi.org/10.1038/d41586-023-02361-7)] [Medline: [37491395](https://pubmed.ncbi.nlm.nih.gov/37491395/)]
14. Langevin M, Grebner C, Güssregen S, Sauer S, Li Y, Matter H, et al. Impact of applicability domains to generative artificial intelligence. *ACS Omega* 2023 Jun 27;8(25):23148-23167 [[FREE Full text](#)] [doi: [10.1021/acsomega.3c00883](https://doi.org/10.1021/acsomega.3c00883)] [Medline: [37396211](https://pubmed.ncbi.nlm.nih.gov/37396211/)]
15. Park JE, Vollmuth P, Kim N, Kim HS. Research highlight: use of generative images created with artificial intelligence for brain tumor imaging. *Korean J Radiol* 2022 May;23(5):500-504 [[FREE Full text](#)] [doi: [10.3348/kjr.2022.0033](https://doi.org/10.3348/kjr.2022.0033)] [Medline: [35434978](https://pubmed.ncbi.nlm.nih.gov/35434978/)]
16. Biswas SS. Role of chat GPT in public health. *Ann Biomed Eng* 2023 May;51(5):868-869. [doi: [10.1007/s10439-023-03172-7](https://doi.org/10.1007/s10439-023-03172-7)] [Medline: [36920578](https://pubmed.ncbi.nlm.nih.gov/36920578/)]
17. van Dis EA, Bollen J, Zuidema W, van Rooij R, Bockting CL. ChatGPT: five priorities for research. *Nature* 2023 Feb;614(7947):224-226. [doi: [10.1038/d41586-023-00288-7](https://doi.org/10.1038/d41586-023-00288-7)] [Medline: [36737653](https://pubmed.ncbi.nlm.nih.gov/36737653/)]
18. Leech NL, Onwuegbuzie AJ. An array of qualitative data analysis tools: a call for data analysis triangulation. *School Psychol Q* 2007 Dec;22(4):557-584. [doi: [10.1037/1045-3830.22.4.557](https://doi.org/10.1037/1045-3830.22.4.557)]
19. Firmin RL, Bonfils KA, Luther L, Minor KS, Salyers MP. Using text-analysis computer software and thematic analysis on the same qualitative data: a case example. *Qual Psychol* 2017 Nov;4(3):201-210. [doi: [10.1037/qup0000050](https://doi.org/10.1037/qup0000050)]
20. Hamilton L, Elliott D, Quick A, Smith S, Choplin V. Exploring the use of AI in qualitative analysis: a comparative study of guaranteed income data. *Int J Qual Method* 2023 Sep 07;22. [doi: [10.1177/16094069231201504](https://doi.org/10.1177/16094069231201504)]
21. Noble H, Smith J. Issues of validity and reliability in qualitative research. *Evid Based Nurs* 2015 Apr;18(2):34-35 [[FREE Full text](#)] [doi: [10.1136/eb-2015-102054](https://doi.org/10.1136/eb-2015-102054)] [Medline: [25653237](https://pubmed.ncbi.nlm.nih.gov/25653237/)]
22. Moore DJ, Pasipanodya EC, Umlauf A, Rooney AS, Gouaux B, Depp CA, et al. Individualized texting for adherence building (iTAB) for methamphetamine users living with HIV: a pilot randomized clinical trial. *Drug Alcohol Depend* 2018 Aug 01;189:154-160 [[FREE Full text](#)] [doi: [10.1016/j.drugalcdep.2018.05.013](https://doi.org/10.1016/j.drugalcdep.2018.05.013)] [Medline: [29958127](https://pubmed.ncbi.nlm.nih.gov/29958127/)]
23. Becker MH. The health belief model and personal health behavior. *Health Educ Monogr* 1974;2:324-508.
24. Ajzen I. The theory of planned behavior. *Organ Behav Hum Decis Process* 1991 Dec;50(2):179-211. [doi: [10.1016/0749-5978\(91\)90020-T](https://doi.org/10.1016/0749-5978(91)90020-T)]
25. Bandura A. *Social Foundations of Thought and Action: A Social Cognitive Theory*. Hoboken, NJ: Prentice Hall; 1986.
26. de Vries H, Dijkstra M, Kuhlman P. Self-efficacy: the third factor besides attitude and subjective norm as a predictor of behavioural intentions. *Health Educ Res* 1988;3(3):273-282. [doi: [10.1093/her/3.3.273](https://doi.org/10.1093/her/3.3.273)]
27. Montoya JL, Georges S, Poquette A, Depp CA, Atkinson JH, Moore DJ. Refining a personalized mHealth intervention to promote medication adherence among HIV+ methamphetamine users. *AIDS Care* 2014;26(12):1477-1481 [[FREE Full text](#)] [doi: [10.1080/09540121.2014.924213](https://doi.org/10.1080/09540121.2014.924213)] [Medline: [24911433](https://pubmed.ncbi.nlm.nih.gov/24911433/)]
28. Knox WB, Stone P. Augmenting reinforcement learning with human feedback. In: *Proceedings of the ICML Workshop on New Developments in Imitation Learning*. 2011 Presented at: ICML 2011; June 28-July 2, 2011; Bellevue, WA URL: <https://www.ias.informatik.tu-darmstadt.de/uploads/Research/ICML2011/icml11il-knox.pdf>
29. Meskó B. Prompt engineering as an important emerging skill for medical professionals: tutorial. *J Med Internet Res* 2023 Oct 04;25:e50638 [[FREE Full text](#)] [doi: [10.2196/50638](https://doi.org/10.2196/50638)] [Medline: [37792434](https://pubmed.ncbi.nlm.nih.gov/37792434/)]
30. Miles MB, Huberman AM. *Qualitative Data Analysis: An Expanded Sourcebook*. Thousand Oaks, CA: SAGE Publications; 1994.
31. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics* 1977 Mar;33(1):159-174. [Medline: [843571](https://pubmed.ncbi.nlm.nih.gov/843571/)]
32. Fu Z, Hsu YC, Chan CS, Lau CM, Liu J, Yip PS. Efficacy of ChatGPT in Cantonese sentiment analysis: comparative study. *J Med Internet Res* 2024 Jan 30;26:e51069 [[FREE Full text](#)] [doi: [10.2196/51069](https://doi.org/10.2196/51069)] [Medline: [38289662](https://pubmed.ncbi.nlm.nih.gov/38289662/)]
33. Xiao Z, Yuan X, Liao QV, Abdelghani R, Oudeyer PY. Supporting qualitative analysis with large language models: combining codebook with GPT-3 for deductive coding. In: *Proceedings of the 28th International Conference on Intelligent User Interfaces*. 2023 Presented at: IUI '23 Companion; March 27-31, 2023; Sydney, Australia URL: <https://dl.acm.org/doi/10.1145/3581754.3584136> [doi: [10.1145/3581754.3584136](https://doi.org/10.1145/3581754.3584136)]
34. Hruschka DJ, Schwartz D, St.John DC, Picone-Decaro E, Jenkins RA, Carey JW. Reliability in coding open-ended data: lessons learned from HIV behavioral research. *Field Methods* 2016 Jul 24;16(3):307-331. [doi: [10.1177/1525822x04266540](https://doi.org/10.1177/1525822x04266540)]
35. Roberts K, Dowell A, Nie JB. Attempting rigour and replicability in thematic analysis of qualitative research data; a case study of codebook development. *BMC Med Res Methodol* 2019 Mar 28;19(1):66 [[FREE Full text](#)] [doi: [10.1186/s12874-019-0707-y](https://doi.org/10.1186/s12874-019-0707-y)] [Medline: [30922220](https://pubmed.ncbi.nlm.nih.gov/30922220/)]
36. O'Connor C, Joffe H. Intercoder reliability in qualitative research: debates and practical guidelines. *Int J Qual Method* 2020 Jan 22;19. [doi: [10.1177/1609406919899220](https://doi.org/10.1177/1609406919899220)]

37. Smith JA. Qualitative Psychology: A Practical Guide to Research Methods. Thousand Oaks, CA: SAGE Publications; 2003.
38. MacPhail C, Khoza N, Abler L, Ranganathan M. Process guidelines for establishing Inter-coder Reliability in qualitative studies. *Qual Res* 2015 Apr 20;16(2):198-212. [doi: [10.1177/1468794115577012](https://doi.org/10.1177/1468794115577012)]
39. Riley WT, Glasgow RE, Etheredge L, Abernethy AP. Rapid, responsive, relevant (R3) research: a call for a rapid learning health research enterprise. *Clin Transl Med* 2013 May 10;2(1):10 [FREE Full text] [doi: [10.1186/2001-1326-2-10](https://doi.org/10.1186/2001-1326-2-10)] [Medline: [23663660](https://pubmed.ncbi.nlm.nih.gov/23663660/)]
40. Timimi F, Ray S, Jones E, Aase L, Hoffman K. Patient-reported outcomes in online communications on statins, memory, and cognition: qualitative analysis using online communities. *J Med Internet Res* 2019 Nov 28;21(11):e14809 [FREE Full text] [doi: [10.2196/14809](https://doi.org/10.2196/14809)] [Medline: [31778117](https://pubmed.ncbi.nlm.nih.gov/31778117/)]
41. Skeen SJ, Jones SS, Cruse CM, Horvath KJ. Integrating natural language processing and interpretive thematic analyses to gain human-centered design insights on HIV mobile health: proof-of-concept analysis. *JMIR Hum Factors* 2022 Jul 21;9(3):e37350 [FREE Full text] [doi: [10.2196/37350](https://doi.org/10.2196/37350)] [Medline: [35862171](https://pubmed.ncbi.nlm.nih.gov/35862171/)]
42. Leeson W, Resnick A, Alexander D, Rovers J. Natural language processing (NLP) in qualitative public health research: a proof of concept study. *Int J Qual Method* 2019 Nov 13;18. [doi: [10.1177/1609406919887021](https://doi.org/10.1177/1609406919887021)]
43. Guetterman TC, Chang T, DeJonckheere M, Basu T, Scruggs E, Vydiswaran VG. Augmenting qualitative text analysis with natural language processing: methodological study. *J Med Internet Res* 2018 Jun 29;20(6):e231 [FREE Full text] [doi: [10.2196/jmir.9702](https://doi.org/10.2196/jmir.9702)] [Medline: [29959110](https://pubmed.ncbi.nlm.nih.gov/29959110/)]
44. Ruksakulpiwat S, Kumar A, Ajibade A. Using ChatGPT in medical research: current status and future directions. *J Multidiscip Healthc* 2023 May 30;16:1513-1520 [FREE Full text] [doi: [10.2147/JMDH.S413470](https://doi.org/10.2147/JMDH.S413470)] [Medline: [37274428](https://pubmed.ncbi.nlm.nih.gov/37274428/)]
45. Zohny H, McMillan J, King M. Ethics of generative AI. *J Med Ethics* 2023 Feb;49(2):79-80. [doi: [10.1136/jme-2023-108909](https://doi.org/10.1136/jme-2023-108909)] [Medline: [36693706](https://pubmed.ncbi.nlm.nih.gov/36693706/)]
46. Thirunavukarasu AJ, Ting DS, Elangovan K, Gutierrez L, Tan TF, Ting DS. Large language models in medicine. *Nat Med* 2023 Aug;29(8):1930-1940. [doi: [10.1038/s41591-023-02448-8](https://doi.org/10.1038/s41591-023-02448-8)] [Medline: [37460753](https://pubmed.ncbi.nlm.nih.gov/37460753/)]
47. Meskó B, Topol EJ. The imperative for regulatory oversight of large language models (or generative AI) in healthcare. *NPJ Digit Med* 2023 Jul 06;6(1):120 [FREE Full text] [doi: [10.1038/s41746-023-00873-0](https://doi.org/10.1038/s41746-023-00873-0)] [Medline: [37414860](https://pubmed.ncbi.nlm.nih.gov/37414860/)]
48. Morley J, DeVito NJ, Zhang J. Generative AI for medical research. *BMJ* 2023 Jul 12;382:1551. [doi: [10.1136/bmj.p1551](https://doi.org/10.1136/bmj.p1551)] [Medline: [37437947](https://pubmed.ncbi.nlm.nih.gov/37437947/)]
49. Thapa S, Adhikari S. ChatGPT, bard, and large language models for biomedical research: opportunities and pitfalls. *Ann Biomed Eng* 2023 Dec 16;51(12):2647-2651. [doi: [10.1007/s10439-023-03284-0](https://doi.org/10.1007/s10439-023-03284-0)] [Medline: [37328703](https://pubmed.ncbi.nlm.nih.gov/37328703/)]
50. Nowell LS, Norris JM, White DE, Moules NJ. Thematic analysis. *Int J Qual Method* 2017 Oct 02;16(1). [doi: [10.1177/1609406917733847](https://doi.org/10.1177/1609406917733847)]

Abbreviations

- AI:** artificial intelligence
- GenAI:** generative artificial intelligence
- ICR:** intercoder reliability
- iTAB:** individualized texting for adherence building
- LLM:** large language model
- NLP:** natural language processing

Edited by K El Emam, B Malin; submitted 11.11.23; peer-reviewed by L Hamilton, S De Paoli, S Biswas; comments to author 29.02.24; revised version received 25.03.24; accepted 06.06.24; published 02.08.24.

Please cite as:

Prescott MR, Yeager S, Ham L, Rivera Saldana CD, Serrano V, Narez J, Paltin D, Delgado J, Moore DJ, Montoya J

Comparing the Efficacy and Efficiency of Human and Generative AI: Qualitative Thematic Analyses

JMIR AI 2024;3:e54482

URL: <https://ai.jmir.org/2024/1/e54482>

doi: [10.2196/54482](https://doi.org/10.2196/54482)

PMID:

©Maximo R Prescott, Samantha Yeager, Lillian Ham, Carlos D Rivera Saldana, Vanessa Serrano, Joey Narez, Dafna Paltin, Jorge Delgado, David J Moore, Jessica Montoya. Originally published in JMIR AI (<https://ai.jmir.org>), 02.08.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium,

provided the original work, first published in JMIR AI, is properly cited. The complete bibliographic information, a link to the original publication on <https://www.ai.jmir.org/>, as well as this copyright and license information must be included.

Original Paper

Reidentification of Participants in Shared Clinical Data Sets: Experimental Study

Daniela Wiepert¹, BA; Bradley A Malin^{2,3,4}, PhD; Joseph R Duffy¹, PhD; Rene L Utianski¹, PhD; John L Stricker¹, PhD; David T Jones¹, MD; Hugo Botha¹, MBChB

¹Department of Neurology, Mayo Clinic, Rochester, MN, United States

²Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN, United States

³Department of Biostatistics, Vanderbilt University Medical Center, Nashville, TN, United States

⁴Department of Computer Science, Vanderbilt University Medical Center, Nashville, TN, United States

Corresponding Author:

Hugo Botha, MBChB

Department of Neurology

Mayo Clinic

200 1st St

Rochester, MN, 55905

United States

Phone: 1 5072841588

Email: botha.hugo@mayo.edu

Abstract

Background: Large curated data sets are required to leverage speech-based tools in health care. These are costly to produce, resulting in increased interest in data sharing. As speech can potentially identify speakers (ie, voiceprints), sharing recordings raises privacy concerns. This is especially relevant when working with patient data protected under the Health Insurance Portability and Accountability Act.

Objective: We aimed to determine the reidentification risk for speech recordings, without reference to demographics or metadata, in clinical data sets considering both the size of the *search space* (ie, the number of comparisons that must be considered when reidentifying) and the nature of the speech recording (ie, the type of speech task).

Methods: Using a state-of-the-art speaker identification model, we modeled an adversarial attack scenario in which an adversary uses a large data set of identified speech (hereafter, the *known set*) to reidentify as many unknown speakers in a shared data set (hereafter, the *unknown set*) as possible. We first considered the effect of search space size by attempting reidentification with various sizes of known and unknown sets using VoxCeleb, a data set with recordings of natural, connected speech from >7000 healthy speakers. We then repeated these tests with different types of recordings in each set to examine whether the nature of a speech recording influences reidentification risk. For these tests, we used our clinical data set composed of recordings of elicited speech tasks from 941 speakers.

Results: We found that the risk was inversely related to the number of comparisons an adversary must consider (ie, the search space), with a positive linear correlation between the number of false acceptances (FAs) and the number of comparisons ($r=0.69$; $P<.001$). The true acceptances (TAs) stayed relatively stable, and the ratio between FAs and TAs rose from 0.02 at 1×10^5 comparisons to 1.41 at 6×10^6 comparisons, with a near 1:1 ratio at the midpoint of 3×10^6 comparisons. In effect, risk was high for a small search space but dropped as the search space grew. We also found that the nature of a speech recording influenced reidentification risk, with nonconnected speech (eg, vowel prolongation: FA/TA=98.5; alternating motion rate: FA/TA=8) being harder to identify than connected speech (eg, sentence repetition: FA/TA=0.54) in cross-task conditions. The inverse was mostly true in within-task conditions, with the FA/TA ratio for vowel prolongation and alternating motion rate dropping to 0.39 and 1.17, respectively.

Conclusions: Our findings suggest that speaker identification models can be used to reidentify participants in specific circumstances, but in practice, the reidentification risk appears small. The variation in risk due to search space size and type of speech task provides actionable recommendations to further increase participant privacy and considerations for policy regarding public release of speech recordings.

KEYWORDS

reidentification; privacy; adversarial attack; health care; speech disorders; voiceprint

Introduction

Background

Advances in machine learning and acoustic signal processing, along with widely available analysis software and computational resources, have resulted in an increase in voice- and speech-based (hereafter referred to as speech for simplicity) diagnostic and prognostic tools in health care [1]. Applications of such technology range from the early detection of cardiovascular [2], respiratory [3], and neurological [4] diseases to the prediction of disease severity [5] and evaluation of response to treatment [6]. These advances have substantial potential to enhance patient care within neurology given the global burden of neurological diseases [7,8], the poor global access to neurological expertise [9,10], and the established role of speech examination within the fields of neurology and speech-language pathology [11].

Large curated data sets are needed to harness the advances in this area. These data sets are costly to assemble and require rare domain expertise to annotate, leading to increased interest in data sharing among investigators and industry partners. However, given the potentially identifiable nature of voice or speech recordings and the health information contained within such recordings, significant privacy concerns emerge. For many data sets, conventional deidentification approaches that remove identifying metadata (eg, participant demographics and date and location of recording) are sufficient, but sharing speech recordings comes with additional risk as the speech signal itself has the potential to act as a personal identifier [12-14]. In recognition of this potential problem, voiceprints are specifically mentioned as an example of biometric identifiers with respect to the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule [15,16]. Approaches that involve modifying nonlinguistic aspects of speech through distortion or alteration of the signal may address the inherent identifiability of the speech signal (ie, its potential as a voiceprint) [13,17], but this is not an option when a central part of speech examination in medicine is to use the acoustic signal to detect subtle nonlinguistic abnormalities indicative of the presence of neurological disease [11,13]. Deidentification in compliance with HIPAA may still be possible under the Expert Determination implementation, whereby the risk of reidentification for unmodified speech recordings is deemed low according to accepted statistical and scientific principles [15,16]. In this respect, various previous studies have investigated the risk of reidentification in research cohort data sets based on demographic or other metadata that may link a participant to their corresponding recordings [18-20], but none have explicitly assessed the inherent risk of the acoustic signal itself. Determining the risk of reidentification for recordings in speech data sets and learning how to best mitigate such risk is necessary for health care institutions to protect patients, research participants, and themselves.

Unfortunately, the same machine learning advances that facilitate the use of speech in health care have also made adversarial attacks, such as deanonymization or reidentification attacks, more feasible. For example, attempting to reidentify a speaker from only a speech recording relies on the mature, well-researched field of speaker identification [21,22]. Studies using speaker identification suggest that the potential for identification from the acoustic signal alone is high [23], although there have been minimal studies in the context of adversarial attacks that may result in potential harm to a speaker [24,25]. Only one previous study has relied on a speaker identification model for reidentification, and the results suggested that the risk was high with a single unknown or unidentified speaker and a moderately small reference set of 250 known or identified speakers [25]. As such, the risk inherent in the acoustic signal, devoid of metadata, is nonzero but relatively unknown, and the feasibility for larger data sets is unexplored.

In addition, these approaches are rarely applied to medical speech data sets [26]. This presents a gap in research as medical speech recordings differ from speech recordings of healthy speakers in a few systematic ways. First, the recordings typically contain speech with abnormalities (ie, speech disorders), which may make reidentification harder as many speech disorders are the result of progressive neurological disease, which causes changes in speech that evolve over months to years [11]. Matching recordings from a time when a speaker was healthy or mildly affected to recordings in which they have a more severe speech disorder may be more difficult [27-29]. Second, the premise of speaker identification is that there are recognizable between-speaker differences tied to identity. However, in a cohort enriched with speech with abnormalities, a substantial proportion of the variance would be tied to the underlying speech disorder as this causes recognizable deviations [11], resulting in speakers sounding less distinct [30]. Finally, medical speech recordings typically contain responses to elicited speech tasks rather than the unstructured connected speech typically used in identification experiments. Some speech task responses do contain connected speech (eg, paragraph reading), but others are very dissimilar (eg, vowel prolongation). The impact of speech task on identifiability remains unknown.

Objectives

In this study, we addressed the risk of reidentification in a series of experiments exploring the reidentifiability of medical speech recordings without using any metadata. We accomplished this goal by modeling an adversarial attack using a state-of-the-art speaker identification architecture wherein an adversary trains the speaker identification model on publicly available, identified recordings and applies the model to a set of unidentified clinical recordings.

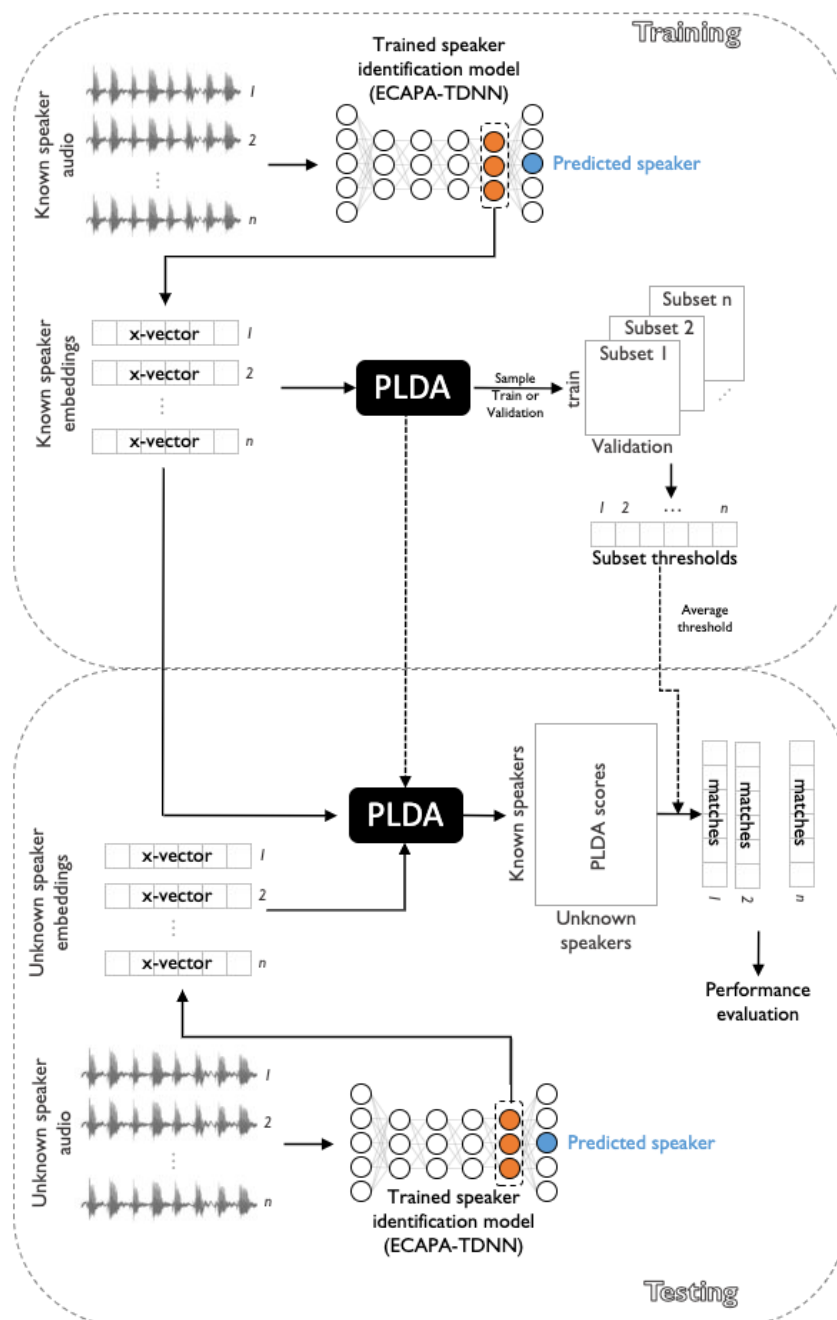
Methods

Overview

Our experimental design was based on the following assumptions: (1) a data recipient has decided to attempt reidentification of study participant data, thereby becoming an adversary; and (2) this adversary relies on an adversarial attack strategy known as a marketer attack, wherein they use a large data set of identified speech (hereafter referred to as the known set), perhaps obtained from a web source such as YouTube, to train a speaker identification model that is then used to reidentify

as many unknown speakers in the shared clinical data set (hereafter referred to as the unknown set) as possible [19,31]. Other attack scenarios are possible, but a marketer attack establishes an accepted baseline for risk. To simulate this attack scenario, we built a text-independent speaker identification model with a combination of x-vector extraction using Emphasized Channel Attention, Propagation, and Aggregation in Time-Delay Neural Network (ECAPA-TDNN) [32] and a downstream probabilistic linear discriminant analysis (PLDA)-based classifier [33,34], as described in detail in the following sections. Figure 1 shows the architecture of our model.

Figure 1. Speaker identification system architecture. During training, recordings from known speakers are fed into a pretrained speaker identification model (ECAPA-TDNN) to extract embeddings. These constitute a low-dimensional, latent representation for each recording that is enriched for speaker-identifying features (x-vectors). We used these x-vectors for known speakers to train a probabilistic linear discriminant analysis (PLDA) classifier and generate an average threshold for acceptance or rejection of a speaker match over several subsets. During testing, the extracted x-vectors are fed into the trained PLDA, and the training threshold is applied, resulting in a set of matches (or no matches) for each recording. ECAPA-TDNN: Emphasized Channel Attention, Propagation, and Aggregation in Time-Delay Neural Network.



Data

Overview

An ideal data set for our attack scenario would consist of (1) a set of elicited speech recordings from tasks typically used in clinical or research speech evaluations and (2) a set of unstructured speech recordings including the same speakers as in item 1 but acquired at a different time and place. This would allow us to directly assess the risk of reidentification of medical recordings by training a model on unstructured connected speech, such as what an adversary may find on the web. Such a data set does not exist. As such, we made use of 2 separate data sets. The first was a combination of the well-known VoxCeleb 1 and 2 data sets, which contain recordings from a web source of >7000 speakers [23,35,36]. The second was a medical speech data set from the Mayo Clinic, which contains recordings of commonly used elicited speech tasks but with fewer speakers.

VoxCeleb

The VoxCeleb 1 and 2 data sets are recent large-scale speaker identification data sets containing speech clips extracted from

celebrity interviews on YouTube [23,35,36]. The utterances are examples of natural, real-world speech recorded under variable conditions from speakers of different ages, accents, and ethnicities. VoxCeleb 1 and 2 have a combined total of 1,281,762 recordings from 7363 speakers.

Mayo Clinic Speech Recordings

The Mayo Clinic clinical speech data set consists of recordings from elicited speech tasks in previously recorded speech assessments. Each speaker has a combination of clips from various tasks commonly used in a clinical speech evaluation, including sentence repetition, word repetition, paragraph reading, alternating motion rates (AMRs), sequential motion rates (SMRs), and vowel prolongation [11]. The clips from speakers vary in recording medium (cassette recording vs DVD), microphone distance, degree of background noise, and presence and severity of motor speech disorder or disorders. There are 19,195 recordings from 941 speakers (the breakdown is presented in Table 1).

Table 1. Breakdown of number of recordings and speakers for each task in the Mayo Clinic clinical speech data set.

	Recordings, n (%)	Speakers, n (%)
Vowel prolongation		
“Aaaaaah”	1734 (9.03)	812 (86.3)
AMR^a		
“Puh,” “tuh,” and “kuh”	3921 (20.43)	777 (82.6)
SMR^b		
“Puh-tuh-kuh”	1049 (5.46)	564 (59.9)
Word repetition		
“Catapult” and “catastrophe”	124 (0.65)	62 ^c (6.6)
Other words	4012 ^d (20.9)	354 ^c (37.6)
Sentence repetition		
“My physician...”	238 (1.24)	222 ^e (23.6)
Other sentences	7505 ^d (39.1)	551 ^e (58.6)
Reading passage		
“You wish to know...”	612 (3.19)	501 (53.2)

^aAMR: alternating motion rate.

^bSMR: sequential motion rate.

^c354 total unique speakers.

^dSamples instead of recordings.

^e551 total unique speakers.

X-Vector Extraction Using ECAPA-TDNN

We generated speaker embeddings using a deep neural network to extract fixed-length embedding vectors (x-vectors) from speech recordings [32,34]. This technique has been shown to outperform previous embedding techniques such as i-vectors [37,38] while offering a competitive performance compared to newer end-to-end deep learning approaches [21,22]. Our

network of choice was the state-of-the-art ECAPA-TDNN model, which was pretrained on a speaker identification task using VoxCeleb 1 and 2 [32]. This model extracts a 192-dimensional x-vector for each speech recording. The model is publicly available through SpeechBrain, an open-source artificial intelligence speech toolkit [39], and is hosted on Hugging Face.

PLDA Back-End Classifier

PLDA classifiers are a standard approach for speaker identification due to their ability to reliably extract speaker-specific information from an embedding space using both within- and between-speaker variance [33,40]. PLDA is a dimensionality reduction technique that projects data to a lower-dimensional space where different classes are maximally separated (ie, maximal between-class covariance). The advantage of PLDA over the standard linear discriminant analysis is that it can be generalized to unseen cases [41]. PLDA can then be used to determine whether 2 data points belong to the same class by projecting 2 data points to the latent space and using the distance between them as a measure of similarity. This works well for speaker identification as speaker embeddings are typically fed into a classifier in pairs, where the classifier's role is to optimally reject or accept the hypothesis that the 2 recordings are from the same speaker. PLDA typically uses the log-likelihood ratio (probability of recordings belonging to the same class vs different classes) to measure similarity, commonly referred to as PLDA scores. During training of a PLDA classifier, PLDA scores for each pairwise comparison in the training set are computed and then used to set a threshold for determining potential speaker matches [33,40].

Our classifier was built and trained on a set of x-vectors extracted from either VoxCeleb or Mayo Clinic speech recordings using ECAPA-TDNN functions from SpeechBrain [39]. We aimed to maximize performance by giving the model multiple speech embeddings per speaker during training, each extracted from recordings under different degradation conditions (eg, varying background noise and microphone distances), which were then averaged to create a single speaker embedding [33].

Threshold Calculation for Acceptance or Rejection

During training, an optimal threshold needs to be determined to classify whether a given PLDA score represents a match, which can then be applied to new, unseen recordings. Matches that pass the threshold are then considered accepted matches. Generally, the equal error rate (EER) is used to select the threshold [21,22,24,33,34]. The use of the EER assumes that the cost of a false acceptance (FA) is the same as a false rejection (FR) such that the optimal threshold is 1, where the FA rate (FAR) equals the FR rate [22]. While this may be feasible for smaller data sets, when there are several million comparisons, the EER often generates many potential matches per speaker. As such, this can overwhelm the model early on and make it difficult for an adversary to find reliable matches. To scale up to large numbers of comparisons, the adversary must make decisions on how to calibrate the threshold calculation, such as penalizing FAs more heavily even if some true acceptances (TAs) are missed. From an adversary's perspective, it is less costly to miss TAs if the identified accepted cases have a high likelihood of being true. In effect, precision is more important than recall. The detection cost function (equation 1 [42]) captures this well:

$$\text{minDCF} = C_{FR} \times FR \times \text{prior}_{\text{target}} + C_{FA} \times FA \times (1 - \text{prior}_{\text{target}}) \quad (1)$$

We take the cost of an FR (C_{FR}) multiplied by the total number of FRs and the prior probability of the target and add it to the cost of an FA (C_{FA}) multiplied by the total number of FAs and the complement of the prior probability.

Using this function, a threshold can be found by setting optimal cost and previous terms based on the adversary's perspective (ie, avoiding FAs more aggressively) and then finding the FA and FR values that minimize the detection cost function (minDCF) [42]. For example, as the prior probability of the target is lowered (ie, if an adversary expects a small overlap), the calculation puts more emphasis on avoiding FAs (lower FAR) as compared to the EER. Increasing the cost of FAs and decreasing the cost of FRs further prevents FAs.

We used the minDCF with two parameter configurations: (1) the default configuration for the SpeechBrain implementation of the minDCF, where FAs and FRs are penalized equally ($C_{FA}=1$; $C_{FR}=1$; prior=0.01) [39]; and (2) a strict configuration with a higher penalty for FAs ($C_{FA}=10$; $C_{FR}=0.1$; prior=0.001).

Due to the large amount of training data in VoxCeleb, it was not computationally feasible to select a threshold for the entire set of identified speakers at once. In addition, we wanted to estimate thresholds that were representative of the population rather than any one subset of speakers. We used a bootstrap sampling technique in which we calculated a minDCF threshold on subsets of training speakers and averaged across runs to estimate the optimal threshold. For each run, latent representations from 2 random subsets of 100 speakers were selected from the training data and fed to the minDCF to calculate a threshold. If the 2 subsets had no overlapping speakers, the entire run was discarded as a threshold could not be calculated. We ran this process between 100 and 500 times depending on the overall number of speakers used for training the PLDA. Training with fewer speakers required fewer runs to converge on an optimal threshold.

Generating Experimental Speaker Sets

To model the attack scenario, we randomly sampled our data sets to generate the following speaker subsets:

1. *Known set*: this set represents speakers with identified audio data from a web source that the adversary has access to.
2. *Unknown-only set*: this set represents speakers in a shared data set who do not have identifiable audio on the web. No unknown-only speakers are present in the known set.
3. *Overlap set*: this set is a proxy for speakers in a shared data set who do have identifiable audio somewhere on the web. Some speakers from the known set are randomly selected to create this set.
4. *Unknown set*: this represents the full shared data set, consisting of both the unknown-only set and the overlap set.

The number of speakers per set varied based on the experiment. Furthermore, the number of speech recordings per speaker varied between the known and unknown sets. We used all available speech recordings per speaker in the known set but randomly selected only 1 recording per speaker in the unknown set. For overlapping speakers, the selected recording for the unknown

set was withheld from the known set. The limit of 1 sample per speaker in the unknown set was based on the nature of a supposed real-world data set where all speech is unlinked and partially deidentified, meaning that the adversary needs to separately find potential matches for each recording even if they come from the same speaker.

Because we randomly subsampled speakers to generate these sets, there is variation in the speakers selected for each experiment, which will result in variability in model performance that is dependent only on the data set. To account for this, we generated multiple speaker splits per experiment. The exact number of splits was dependent on the experiment.

Experiments

VoxCeleb Realistic Experiments: Effect of Search Space Size

We relied on VoxCeleb 1 and 2 to investigate the capability of an attack as a function of the size of the search space (ie, the number of comparisons made to find matching speakers). We reidentified speakers by comparing each speaker in the known set to each speaker in the unknown set. Thus, the search space is the product of the sizes of the known and unknown sets. As such, an increase in either set will increase the number of comparisons. We considered both cases separately, which allowed us to consider one scenario that is dependent on the resources of the adversary (known set size) and another that is under the control of the sharing organization (unknown set size).

To construct a realistic scenario, we assumed that the known and unknown sets would have a low degree of speaker overlap. To justify this assumption, one can consider what would be involved in constructing a set of known speakers. In the absence of metadata about the unknown speakers (eg, the ages and location), there would be no way for an adversary to target a specific population to build their known set. It is unlikely to be feasible for an adversary to manually collect and label speech recordings for a large proportion of the population. Instead, an adversary would likely need to rely on a programmatic approach using easily accessible identifiable audio, such as scraping audio from social media and video- or audio-sharing websites [43].

It is worth noting that this would still be difficult because of several confounding factors: (1) not all members of the population use these websites; (2) not all users have publicly accessible accounts; (3) users with publicly accessible accounts may not have identifiable information linked to them; (4) some accounts post audio or video from multiple speakers, including speakers who also have their own accounts; (5) many users do not post at all; and (6) the population of users is not representative of the general US population, let alone the subset with speech disorders—in terms of the distribution of both age and geographic area [44]. As such, there is no reason to suspect that a patient in a shared medical speech data set would have a high likelihood of existing in an adversary's set of identified audio recordings.

We also assumed that the adversary would not know which unknown speakers, if any, exist in the known speaker set. Therefore, the adversary must consider all potential matches rather than only focusing on the N overall best matches, where N is the known overlap. This would reduce the reliability of any match because the likelihood of all potential matches being true is lower than the likelihood of the best N matches being true.

We first trained the speaker identification model with the number of speakers in the known set increasing from 1000 to 7205 while maintaining a static unknown set size of 163 speakers, with low speaker overlap between sets ($n=5$, 3.1% speakers in the overlap set and $n=158$, 96.9% in the unknown-only set).

We then trained the model with a fixed known set size of 6000 speakers while increasing the number of speakers in the unknown set from 150 to 1000 speakers and maintaining a low overlap of 5 speakers.

Given the low number of overlapping speakers and overall large set sizes, we generated 50 speaker splits for each set size of interest (known set: 1000, 4000, and 7205; unknown set: 150, 500, and 1000).

The acceptance threshold for these experiments was set using the strict minDCF configuration. Experimental parameters are summarized in [Table 2](#).

Table 2. Experimental parameters, including number of runs; set sizes; and minimum detection cost function (minDCF) parameters such as the cost of a false acceptance (CFA), cost of a false rejection (CFR), and prior probability (prior).

Experiment	Runs, n	Set size, total speakers				minDCF parameters		
		Known	Unknown	Unknown only	Overlap	C _{FA}	C _{FR}	Prior
VoxCeleb: effect of search space size and known-overlap worst-case scenario	50	<ul style="list-style-type: none"> 1000 to 7205 (varied known) 6000 (varied unknown) 	<ul style="list-style-type: none"> 163 (varied known) 150 to 1000 (varied unknown) 	<ul style="list-style-type: none"> 158 (varied known) 145 to 995 (varied unknown) 	<ul style="list-style-type: none"> 5 	10	0.1	0.001
VoxCeleb: full-overlap worst-case scenario	20	<ul style="list-style-type: none"> 1000 to 7205 (varied known) 6000 (varied unknown) 	<ul style="list-style-type: none"> 163 (varied known) 150 to 1000 (varied unknown) 	<ul style="list-style-type: none"> 0 	<ul style="list-style-type: none"> 163 (varied known) 150 to 1000 (varied unknown) 	10	0.1	0.001
Mayo Clinic speech recordings: cross-task	20	<ul style="list-style-type: none"> 500 	<ul style="list-style-type: none"> 55 	<ul style="list-style-type: none"> 50 	<ul style="list-style-type: none"> 5 	1	1	0.01
Mayo Clinic speech recordings: within task	20	<ul style="list-style-type: none"> 500^a 	<ul style="list-style-type: none"> 55 	<ul style="list-style-type: none"> 50 	<ul style="list-style-type: none"> 5 	1	1	0.01

^aWord repetition: 299 speakers; reading passage: 466 speakers.

VoxCeleb Known-Overlap and Full-Overlap Experiments: Worst-Case Scenarios

There are two important initial assumptions in our construction of realistic experiments: (1) the adversary was unaware of the amount of overlap between known and unknown sets, and (2) the amount of overlap was low. Thus, we considered how reidentification risk would be affected if either assumption was incorrect.

First, we considered a potential worst-case scenario in which the adversary did know the number of overlap speakers N and, therefore, was able to limit potential matches to the top N best matches. As previously mentioned, limiting the number of matches could theoretically improve model reliability, and further reducing the number of matches could produce more noticeable effects. We leveraged our base results from the realistic experiments and only considered the top N best matches.

Next, we considered a less realistic worst-case scenario in which all unknown speakers exist in the known speaker set. From an adversary's perspective, a full-overlap scenario would provide the best chance for them to successfully reidentify speakers because most FAs occur when the model finds a match for unknown speakers who are not in the known speaker set.

We assessed this scenario by replicating the realistic experiments with full overlap between the known and unknown sets. That is, regardless of the unknown set size, all speakers also exist in the known set (no unknown-only set). When increasing the known set size with a fixed unknown set of 163 speakers, the overlap set consists of all 163 speakers, and when increasing the unknown set size with a fixed unknown set, the overlap set is the same as the unknown set size of interest (150, 500, and 1000). In this scenario, we generated only 20 speaker splits for

each set size of interest as the larger overlap set led to less variance across runs.

As in the realistic experiments, the acceptance threshold was set using the strict minDCF configuration. Experimental parameters are summarized in Table 2.

Mayo Clinic Speech Recording Experiments: Effect of Speech Task

Next, we shifted our focus from the public VoxCeleb data set to a private data set of Mayo Clinic medical speech recordings to look at factors specific to a clinical speech data set, such as whether certain elicited tasks are easier for reidentification and whether being able to link recordings to the same speaker across tasks (pooling) increases risk.

We first compared the performance of the speaker identification model across the various elicited speech tasks in the Mayo Clinic data set based on the same adversarial attack scenario used with the VoxCeleb experiments. In this scenario, the cross-task performance aligns with a real-world case in which the training data contain connected speech recordings (ie, recordings of continuous sequences of sounds such as those of spoken language) but speakers are reidentified using a variety of elicited speech tasks (Table 1). Each task has a different degree of similarity to connected speech (left: most; right: least):

Reading passage > sentence repetition > word repetition > SMR > AMR > vowel prolongation

The reading passage is essentially real-world connected speech in terms of content and duration, but sentence repetition is closer to the connected speech seen in most speech data sets [23]. As such, we selected sentence repetition recordings for speakers in the known set.

The resulting known set comprised 500 speakers and included all sentence repetition recordings, excluding any repetitions of

the physician sentence (“My physician wrote out a prescription”), which was saved for the unknown set. We then generated separate unknown sets for each elicited task with 55 speakers ($n=5$, 9% overlap and $n=50$, 91% unknown only) who had both sentence repetition recordings and a recording for the given reidentification task (eg, “My physician...” sentence and AMRs).

The known and unknown set sizes were bounded by the number of speakers with sentence repetition recordings (587 speakers) as the sentence-sentence configuration required enough speakers to create a separate known and unknown-only set. We also considered the sentence-sentence configuration (ie, sentence repetitions in both the known and unknown sets) as the realistic baseline.

As a secondary part of this experiment, we pooled all available recordings from all elicited speech tasks (by averaging their embeddings) to generate an unknown set in which the adversary could link recordings from a given speaker (ie, there would be more speech for each unknown speaker).

In addition to the cross-task performance, we compared the within-task performance—where the same elicited speech task is used for both known and unknown speakers—to determine whether anything about the nature of a given speech task affected reidentification. For example, the variance across recordings for the sentence repetition task reflects a combination of static speaker factors (eg, identity and age), dynamic speaker factors (prosody, eg, the same speaker may emphasize different words in a sentence on repeated trials), and content factors (ie, different words in different sentences). In contrast, a task such as AMR involves repeating the same syllable as regularly and rapidly as possible, with most of the variance across speakers likely resulting from static speaker factors. A priori, considering all the elicited tasks, one would expect the proportion of variance across speakers due to dynamic speaker factors to decrease following the same scale as similarity to natural speech. The reading passage would have the most variance due to dynamic speaker factors alone, whereas vowel prolongation would have the least variance. By removing the confounding variable of different elicited tasks for known and unknown speakers (ie, the model is both trained and tested on the same task), we can ascertain whether the qualities of the speech task itself influence reidentification.

We used the same set sizes as the cross-task experiments (500 known, 55 unknown, and 5 overlap) but used recordings from the same elicited speech task in both the known and unknown sets. This setup required at least 2 recordings per speaker for each task. Some tasks had <500 unique speakers or not enough recordings (word repetition and reading passage), so not every known set had exactly 500 speakers. The word repetition task had 299 speakers, and the reading passage task had 466 speakers.

To account for the decrease in the amount of data as compared to the VoxCeleb experiments, we generated only 20 speaker splits per task with default minDCF parameters. Experimental parameters are summarized in [Table 2](#).

Statistical Analyses

Given that we were simulating an adversarial attack and not optimizing a model, we used random splitting to account for the potential of outlier cases, wherein specific configurations of speakers in the known and unknown sets had a higher-than-average risk of reidentification. We first randomly sampled our larger data set either 20 or 50 times depending on the experiment to generate speaker splits (known, unknown, and overlap sets). We also randomly selected a single recording per speaker in the unknown set to mitigate utterance effects. Furthermore, we used bootstrap sampling of the known (training) set to estimate our acceptance threshold by feeding cohorts of 100 speakers to the minDCF function between 100 and 500 times to converge on an optimal threshold. The exact number of runs was dependent on the overall number of speakers in the known set.

Our primary outcome of interest was the average number of FAs, where the model accepts a match for an unknown speaker without a true match, compared to TAs over several subsampled data sets. Using these counts, we also calculated precision. These metrics informed the reliability of reidentification. Note that TAs and FAs are functionally equivalent to true and false positives, respectively. Using the counts, we also calculated the Pearson correlation coefficient between FAs and set size along with the FAR to determine whether a linear correlation existed between the number of FAs and the number of speakers or comparisons. A 2 tailed t test was performed to determine the significance of each correlation.

Ethical Considerations

The primary data type for this work was clipped speech recordings from either VoxCeleb or our Mayo Clinic clinical speech data set. We could not deidentify the data due to the nature of our work, and the data sets were not anonymous. The VoxCeleb data set has no privacy protections or additional consent processes in place given its public nature—all recordings come from interviews of celebrities posted on YouTube [23,35,36]. For the Mayo Clinic clinical speech data set, we submitted an institutional review board application to the Mayo Clinic to gain permission to use the data. Our work was deemed exempt from additional consent requirements and granted a waiver of HIPAA authorization considering the secondary nature of the analysis. No compensation was offered to participants in the original studies. As the clinical data set may contain private health information, we do not share any recordings or models trained on the clinical recordings. Only researchers at our institution with proper permission can access the clinical data set.

Results

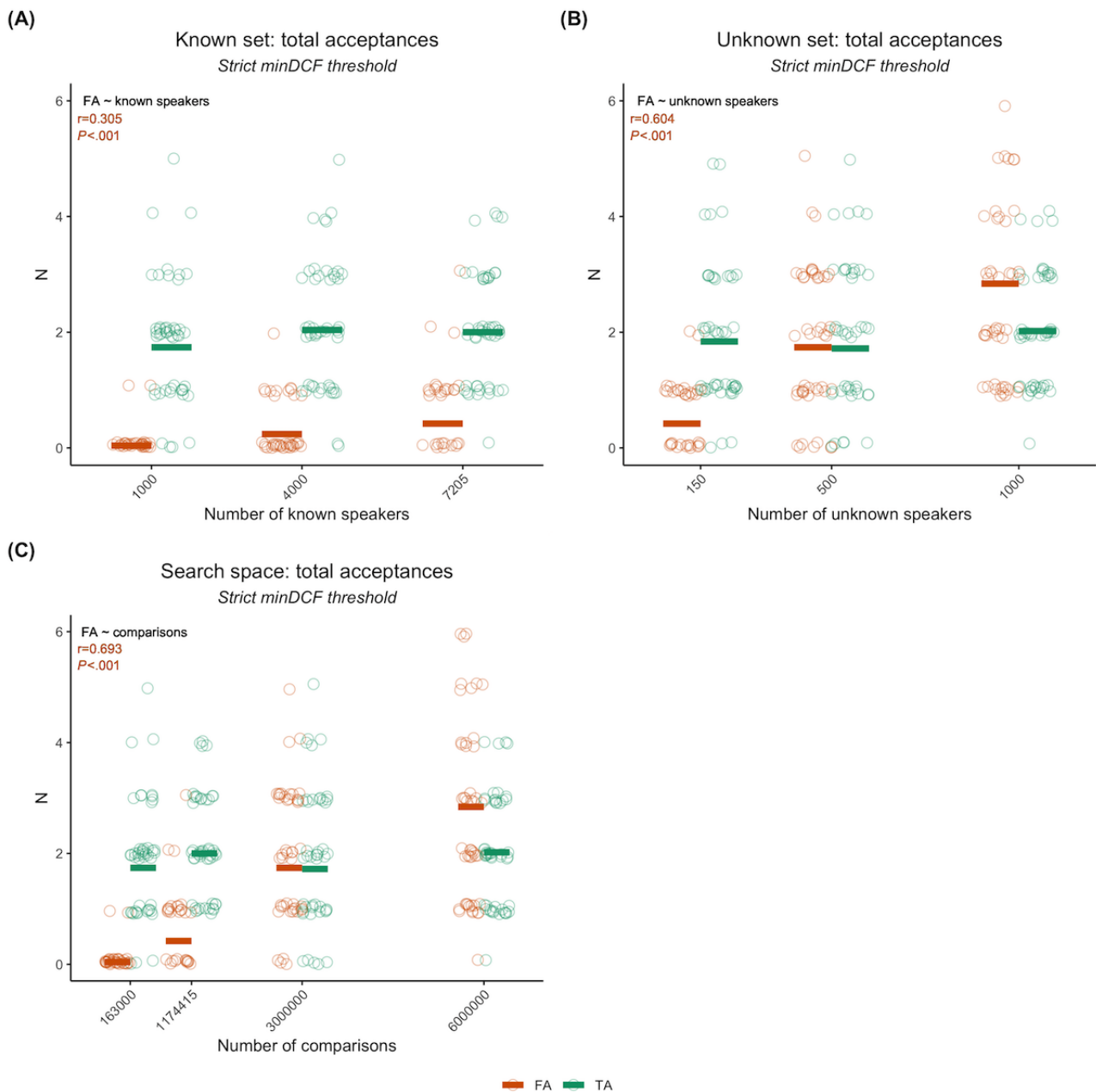
VoxCeleb Realistic Experiments: Effect of Search Space Size

When training the speaker identification model with increasing numbers of speakers in the known set while maintaining a static unknown set size with low speaker overlap between sets, we found that increasing the number of speakers in the known set resulted in an increase in the mean number of FAs while TAs

remained stable, with a linear correlation between FAs and the number of known speakers ($r=0.30$; $P<.001$; $t_{148}=3.89$; Figure 2A). Increasing the size of the unknown set had a similar yet more pronounced effect than increasing the known set size, with a higher linear correlation between FAs and the number of unknown speakers ($r=0.60$; $P<.001$; $t_{148}=9.21$; Figure 2B).

The difference in effect can be understood based on the geometry of the search space. While the unknown set remains substantially smaller than the known set, adding a speaker to the unknown set will result in a larger increase in the search space than adding a speaker to the known set. As such, we can better demonstrate the overall trend in FAs by considering the results in terms of total comparisons (ie, search space size) rather than individual set size.

Figure 2. Number of true acceptances (TAs) and false acceptances (FAs) for the speaker recognition model in a realistic scenario using VoxCeleb. (A) shows the counts when varying the number of known speakers while keeping the number of unknown speakers static, (B) shows the counts when varying the number of unknown speakers while keeping the number of known speakers static, and (C) shows the overall trend in terms of the number of comparisons made (ie, the search space size= $\text{known} \times \text{unknown speakers}$). All plots (A-C) include the Pearson correlation coefficient and corresponding significance for FAs and number of speakers or comparisons. Each run is plotted as a single circle, with red horizontal lines indicating the mean number of FAs and green horizontal lines indicating the mean number of TAs. minDCF: minimum detection cost function.



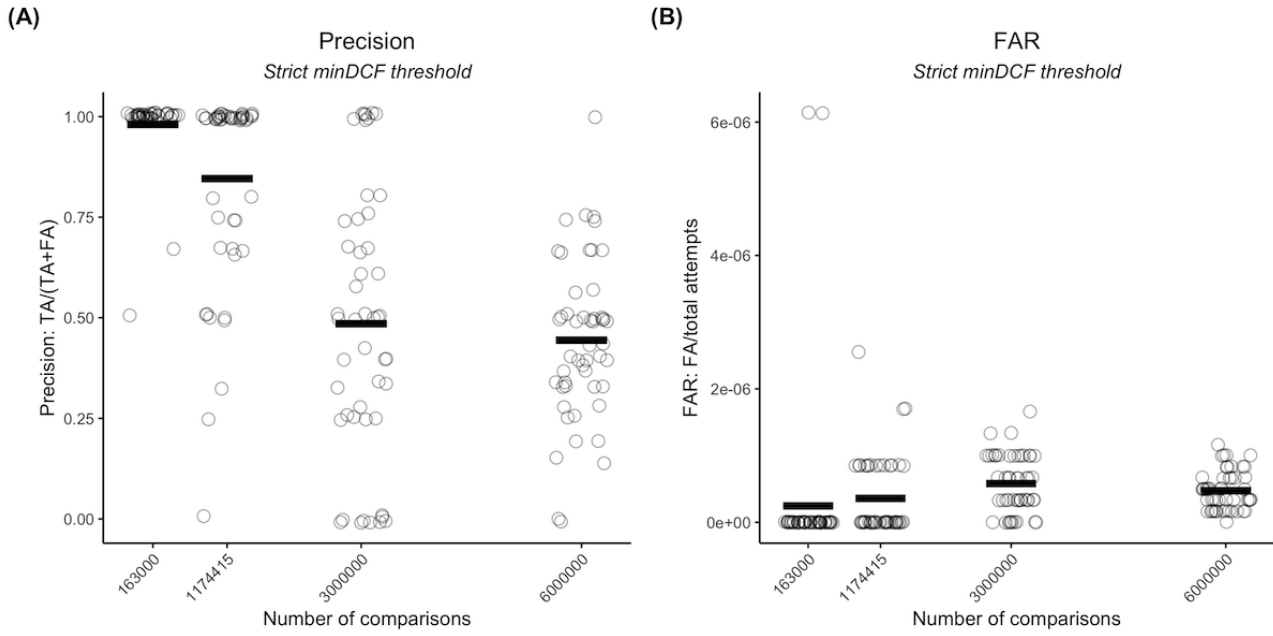
We observed that there was a high positive linear correlation between FAs and the number of comparisons ($r=0.69$; $P<.001$; $t_{198}=13.54$; Figure 2C), with the mean FAs increasing from 0.04 to 2.84 while TAs remained stable. The ratio between FA and

TA (FA/TA) rose from 0.02 at 1×10^5 comparisons to 1.41 at 6×10^6 comparisons, with a near 1:1 ratio at the midpoint of 3×10^6 comparisons. There was a corresponding drop in precision (Figure 3A). It was notable that the FAR remained low and

relatively stable, averaging at 4.152×10^{-7} (SD 7.255×10^{-7} ; Figure 3B), indicating that the demonstrated trend should hold for the larger numbers of comparisons that we would expect to see in a real attack.

We further observed that using a stricter threshold for matches resulted in our model selecting only 1 match per speaker. This is functionally the same as limiting matches to only the best potential match for each speaker (rank-1 matches), which is an option for an adversary to increase reliability without knowledge of the amount of overlap.

Figure 3. Precision and false acceptance rates (FARs) for the speaker recognition model in a realistic scenario using VoxCeleb. Precision (A) and FARs (B) are shown as a function of the number of comparisons. For both plots, each run is represented by a circle, and the mean is represented by a horizontal black line. FA: false acceptance; minDCF: minimum detection cost function; TA: true acceptance.

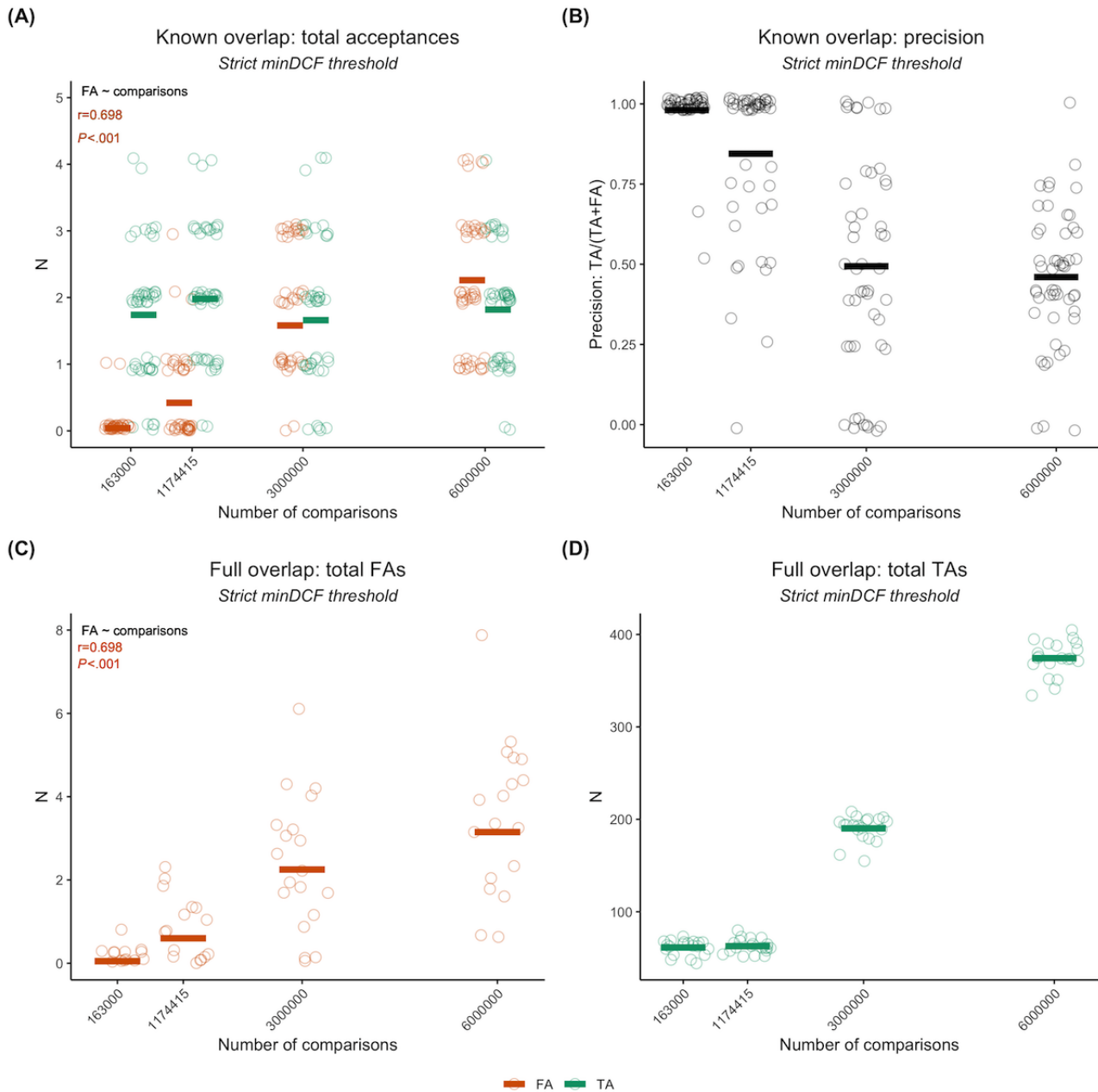


VoxCeleb Known-Overlap and Full-Overlap Experiments: Worst-Case Scenarios

When only considering the top N best matches, we found that there was still a trend of increasing FAs, with a high linear correlation with the number of comparisons ($r=0.70$; $P<.001$; $t_{198}=13.72$; Figure 4A). The FA/TA ratio increased from 0.02 at 1×10^5 comparisons to 1.24 at 6×10^6 comparisons and again had a near 1:1 ratio at 3×10^6 comparisons. These results indicate that some FAs were seen as better matches than some TAs, as further supported by the associated drop in precision (Figure 4B).

When all unknown speakers existed in the known speaker set, the performance improved significantly, with most matches being correct (Figure 4C). Even so, there was still a high positive linear trend for FAs, indicating that, at high overlap, some FAs were ranked higher than TAs ($r=0.67$; $P<.001$; $t_{78}=7.98$; Figure 4D). The FA/TA ratio exhibited a fairly large increase considering the number of TAs, increasing from 0.0008 at 1×10^5 comparisons to 0.008 at 6×10^6 comparisons. This is surprising given that, for the realistic experiments, all FAs were associated with matches for nonoverlapping speakers.

Figure 4. Results for our speaker recognition model in worst-case scenarios using VoxCeleb. (A) shows the true acceptance (TA) and false acceptance (FA) counts for a known-overlap scenario (limited to N=5 best matches), whereas (B) shows the corresponding precision as a function of the number of comparisons (search space size). (C) and (D) show the FA and TA counts for a full-overlap scenario in which all unknown speakers are present in the known speaker set as a function of the number of comparisons (search space size). (A) and (C) also show the Pearson correlation coefficient and corresponding significance between FAs and number of comparisons. Each run is plotted as a single circle, with red horizontal lines indicating the mean number of FAs, green horizontal lines indicating the mean number of TAs, and black horizontal lines indicating the mean precision. minDCF: minimum detection cost function.



Mayo Clinic Speech Recording Experiments: Effect of Speech Task

We first compared the performance of the speaker identification model across the various elicited speech tasks in the Mayo Clinic data set based on the same adversarial attack scenario used in the VoxCeleb experiments. We observed that the total number of acceptances decreased as the unknown speaker tasks became less similar to the known speaker task, but the proportion of TAs and FAs also varied. This made it more difficult to determine the performance through counts alone (Figure 5A). When considering precision and FA/TA ratio instead, we found

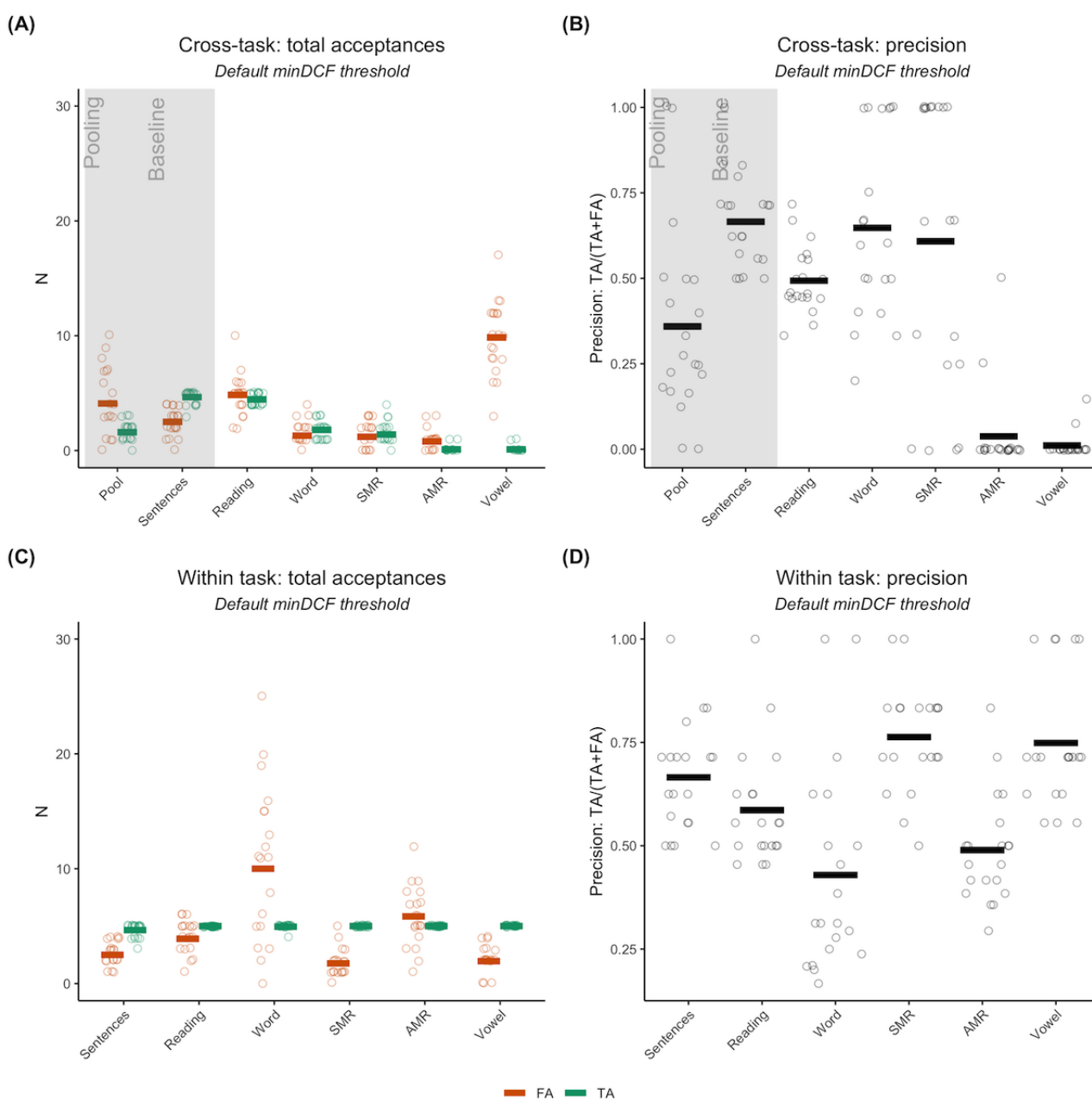
that the baseline (sentence-sentence) had the best performance, although the average precision was not high (FA/TA=0.54; precision=66.5%; Figure 5B). The paragraph reading, word repetition, and SMR tasks had a worse performance than the baseline but were comparable to each other in terms of both precision (Figure 5B) and FA/TA ratios (reading passage: FA/TA=1.09; word repetition: FA/TA=0.72; SMR: FA/TA=0.85). However, the AMR and vowel prolongation tasks had extremely low precision and high FA/TA ratios. Vowel prolongation, in particular, had a precision of 0 (almost no TAs across runs) but a high number of FAs, resulting in a ratio of 98.5. Pooling resulted in decreased performance compared to

the baseline and the top-performing tasks in terms of both precision (approximately 36%) and FA/TA ratio (2.56). This was likely due to the influence of AMR and vowel prolongation recordings.

The within-task results did not exhibit the same effect as the cross-task results. We found that all tasks reidentified the overlapping speakers (TA=10) but the number of FAs varied drastically across tasks (Figure 5C). Previously, the baseline

had the best performance, whereas we instead observed that the SMR and vowel prolongation tasks had the highest precision (Figure 5D), as well as FA/TA ratios of 0.35 and 0.39, respectively. In fact, as tasks became more dissimilar from connected speech and had less variance due to dynamic speaker factors, they saw a relative increase in performance compared to the cross-task scenario. Word repetition was the only exception to this, with lower precision and a greater FA/TA ratio of 2.02 as compared to the cross-task performance.

Figure 5. Results for our speaker recognition model using the Mayo Clinic clinical speech data set. (A) and (B) show cross-task results, in which recordings for known speakers are always sentence repetition but the task for unknown speaker recordings varies. The baseline is when sentence repetitions are in both the known and unknown sets. Pooling is when all recordings for an unknown speaker are linked together across all tasks. (A) shows the breakdown of counts for this case, whereas (B) is the corresponding precision. (C) and (D) show within-task results, where tasks for known and unknown speakers are always the same. (C) is the breakdown of counts for this case, whereas (D) is the corresponding precision. Each run is plotted as a single circle, with red horizontal lines indicating the mean number of false acceptances (FAs), green horizontal lines indicating the mean number of true acceptances (TAs), and black horizontal lines indicating the mean precision. AMR: alternating motion rate; minDCF: minimum detection cost function; SMR: sequential motion rate.



Discussion

Principal Findings

In this study, we investigated the risk of reidentification of unidentified speech recordings without any other speaker- or recording-related metadata. To do so, we performed a series of experiments reflecting a marketer attack by an adversary with access to identified recordings from a large set of speakers and the capability to train a speaker identification model, which would then be used to reidentify unknown speakers in a shared data set. We systematically considered how changes in the size of the data sets and the nature of the speech recordings affected the risk of reidentification. We found that it is feasible to use a speaker identification design—a deep learning speaker embedding extractor (x-vectors) coupled with a PLDA back end—to reidentify speakers in an unknown set of recordings by matching them to recordings from a set of known speakers. Given the performance of current state-of-the-art speaker identification models, this is not surprising. However, these models have only rarely been applied in an adversarial attack scenario [24,25] (ie, their potential as an attack tool for an adversary who aims to reidentify speakers in a shared or publicly available data set was largely unknown). Furthermore, the feasibility of such an attack has not been considered and may have been assumed to be low for speech recordings stripped of all metadata (sometimes referred to as deidentified or anonymous in the literature) without considering the identifiability of the acoustic signal itself [45-48].

Our findings suggest that this is not true. Consistent with a previous study that found a high reidentification risk for an unknown speaker with known sets of up to 250 speakers (search space of ≤ 250 comparisons) [25], we observed that risk was indeed high for small search spaces. For example, when attempting to reidentify 5 overlapping speakers between a small set of unknown speakers ($n=163$) and a moderate set of known speakers ($n=1000$), our model had nearly perfect precision (Figure 3A) and identified 2 speakers on average (FA/TA=0.02; Figure 2A). However, our experiments allowed us to extend this to more realistic search spaces, such as scenarios in which an adversary uses a known speaker set of up to 7205 speakers and an unknown speaker set of up to 1000 speakers (search space of ≤ 6 million comparisons). We observed that the risk dropped sharply as the search space grew. The FAR was relatively stable at 4.152×10^{-7} (Figure 3B), which translates to an average increase of 1 FA for every 2.5 million comparisons. This is a key take-home message from these experiments—increasing the size of the search space, whether by increasing the size of the adversary's set of identified recordings or of the shared data set, resulted in a corresponding increase in the number of FAs. Given that the number of overlapping speakers remained constant, this suggests that the primary driver of FAs is the size of the nonoverlapping known-to-unknown comparison space (ie, most FAs arise from nonoverlapping unknown speakers being falsely matched to known speakers). In fact, all FAs in the realistic experiments corresponded to nonoverlapping unknown speakers. Here, it is worth noting that, in the experiments in which we only considered the top N matches (where N =number of overlapping

speakers), this trend remained true because some of the FAs scored higher than TAs (Figure 4). This suggests that for a sufficiently large search space, even considering only the best N matches will result in many FAs. We pushed this line of reasoning to its limit by considering a worst-case scenario of full overlap in which all unknown speakers had a true match. Even in this scenario, there were still many FAs, and the proportion of FAs increased with increasing search space size. Importantly, this scenario showed that overlapping speakers can still be falsely matched when the overlap is high.

Our experiments with the Mayo Clinic clinical speech recordings allowed us to assess the influence of speech task based on both cross-task and within-task performance. When the model was trained on sentence repetition (ie, the known data set consisted of sentence recordings) and then applied to other tasks (ie, the unknown set consisted of elicited, nonsentence speech), all tasks performed below the baseline, but performance deteriorated most drastically for the less connected speech-like tasks such as AMR and vowel prolongation. These results can be understood with reference to the default minDCF settings, which would penalize FAs and FRs equally. The threshold was chosen using sentence repetition task recordings such that, in most instances, all overlapping speakers were reidentified for unknown sets with connected speech tasks (sentence repetition, paragraph reading, word repetition, and SMRs). The minDCF threshold for these similar tasks resulted in fewer overall acceptances (higher FR rate), but as the tasks diverged from sentence repetition with respect to the degree of connectedness, they were also less likely to be FAs. This suggests that identifiable characteristics learned from training on the sentence repetition task translate well to other connected speech tasks. It also demonstrates the difficulty of choosing a threshold when the tasks in the known set are different from those in the unknown set. Because of the differences within a speaker across tasks, it becomes hard to balance TAs with the flood of FAs as the search space increases. In this instance, a slightly stricter threshold may have been better for the adversary. In contrast, the non-connected speech tasks (AMRs and vowel prolongation) had almost no TAs and a high number of FAs, suggesting that identifiable characteristics from connected speech tasks do not translate to non-connected speech tasks. This is not unexpected given that models perform worse when tested on data that are dissimilar from the training data [49,50]. Following this, we also found that pooling across tasks decreased performance from the baseline. Generally, having more data for a speaker is expected to improve performance, but it is possible that adding recordings of nonsentence tasks to the unknown set hurt performance because the identifiable characteristics are different across tasks and the system is unable to accommodate them. In other words, any helpful characteristics from the connected speech tasks were cancelled out by competing characteristics from the non-connected speech tasks.

In the within-task scenarios, where the known and unknown sets were made up of the same task, the reidentification power for overlapping speakers was better than in the cross-task scenario, but the tasks exhibited vastly different FA rates. In fact, many tasks that were different from connected speech saw improved performance. For example, vowel prolongation, which

is nonconnected and the most perceptually different from sentence repetition, exhibited the worst cross-task performance but the second-best within-task performance. This may be because less connected tasks have fewer interfering dynamic speaker factors such that they isolate well the acoustic features that are tied to identity.

Another important finding is that performance for sentence repetition was much weaker than expected based on the VoxCeleb experiments with a larger number of comparisons. We suspect that this may be due to a combination of factors. First, it may be more difficult to differentiate speakers in an unknown set of elicited recordings in which every speaker utters the same sentence. Second, the clinical recordings were all made by patients referred for a speech examination. Consequently, the resulting cohort contained mostly speech with abnormalities, which may impact the PLDA performance. Third, the Mayo Clinic clinical speech data set is smaller than the VoxCeleb data set in terms of both the number of speakers and the number of recordings per speaker, and the recordings are also shorter in duration. This likely had a negative impact on the training of the PLDA classification back end. It remains unknown whether larger clinical data sets or data sets with more recordings per speaker may yield findings more similar to the VoxCeleb results.

Taken together, our findings suggest that the risk of reidentification for a set of clinical speech recordings devoid of any metadata in an attack scenario such as the one we considered in this study is influenced by (1) the number of comparisons that an adversary must consider, which is a function of the size of both the unknown and known data sets; (2) the similarity between the tasks or recordings in the unknown and known data sets; and (3) the characteristics of the recordings in the unknown data set, such as degree of speaker variance and presence and type of speech disorders. These findings translate to actionable goals for both an adversary and the sharing organization.

Mitigating Privacy Risk

While we assumed that the sharing organization had already reduced risk by stripping recordings of demographic (eg, age or gender) or recording (eg, date or location) metadata, we additionally suggest that reidentification risks could be further reduced by increasing the search space (ie, larger shared data set size) or decreasing the similarity between shared recordings and publicly available recordings (eg, sharing vowel prolongation recordings as long as a publicly available vowel prolongation recording data set does not exist or sharing a larger variety of speech disorder recordings instead of those for a single disorder). Even if the number of overlapping speakers increased with the size of the shared data set, the results from the full-overlap scenario indicate that a model could still have reduced reliability due to an increasing FAR.

In contrast, an adversary can also use this knowledge to enhance their attacks. From their perspective, any additional information that can reduce the search space or increase the similarity between recordings will increase the reliability of speaker matches. This could involve using demographics such as gender, be they shared or predicted by a separate model, to rapidly reduce the number of comparisons. For instance, when the

gender balance is 50:50, comparing unknown male individuals to known male individuals would reduce the number of comparisons by 75% (eg, from 6 million to 1.5 million). The adversary may also seek out publicly available recordings of speech with abnormalities to refine their model or models or reduce the search space based on speech disorders. If social media groups exist where identified users with certain medical or speech disorders post videos or audio, an adversary could restrict their known set to these users. Similarly, research participants and support staff may also influence risk through disclosure of participation. By disclosing participation in a study known to share speech recordings, a participant would effectively reduce the size of the known set to 1, increasing their individual risk of reidentification. In addition, having a confirmed match can increase risk overall as the adversary would have a baseline to determine the reliability of matches [51]. Although the focus of this investigation was on the change in relative risk with changes in data set size and speech task, it is worth considering our findings in the context of other factors that impact risk in practice. The most obvious factor is the availability of additional metadata on the speakers or recording. In this respect, it is worth noting that sufficient demographic data, even in the absence of speech, are well known to carry a significant risk of reidentification [19,52]. If any aspect of the metadata makes a patient population unique (ie, there is only one person in a given age range), the risk of reidentification increases [12,14]. Furthermore, the risk is not necessarily the same for all speakers or groups. For example, individuals with rare speech disorders, accents, or other qualities may be easier to match across known and unknown data sets. There may also be identifiable content in the recordings. During less structured speech tasks such as recordings of open-ended conversations, participants may disclose identifiable information about themselves (eg, participants saying where they live). Removing these spoken identifiers is an active area of research [25].

However, it is important to acknowledge that simply because records are vulnerable to reidentification does not mean that they would be reidentified. Notably, when assessing privacy concerns, the probability of reidentification during an attack is conditional on the probability of an attack occurring in the first place [52]. In most instances in which data are shared, the receiving organization or individual will not have any incentive to attempt reidentification. The sharing organization and, in some cases, a receiving organization may also take steps to discourage the risk of an attack. These may take the form of legal (eg, data-sharing agreements) or technical (eg, limited, monitored access) deterrents to a reidentification attack [53]. In contrast, the risk of an attack may be higher for publicly available data sets [54], but there may also be a greater risk of reidentification without a targeted attack. For example, in the field of facial recognition, some companies have scraped billions of photos from publicly available websites to create massive databases with tens of millions of unique faces. These are then used to train a matching algorithm [43], which an end user could query using a photo of an unknown face and obtain a ranked list of matching faces and the source (eg, Facebook). The end user can visit the source website and instantly gain access to other data that may increase or decrease their confidence in a match as well as provide feedback on matches, thereby gradually

increasing the performance of the tool as well as the number of known faces. If similar databases are built for speech recordings, they will certainly include publicly available medical speech recordings. Every query to the model would then represent a threat to such a public sample being matched to a queried recording regardless of the intent of the user who queried the model. Such a scenario is difficult to simulate because of the continuously improving nature of the algorithm and the fact that users would incorporate various degrees of nonspeech data.

Refraining from publicly releasing data sets is an obvious mitigation strategy for some of these threats. However, the risk of reidentification must always be balanced with the benefit of data sharing as larger, more representative data sets for the development and testing of digital tools may benefit patients. It is critical that policy makers consider this balance in the context of the rapidly evolving field of artificial intelligence. Naïve approaches such as the “deidentification release-and-forget model” are unlikely to provide sufficient protection [55]. Similarly, informed consent for public release is problematic because the risk of reidentification will be neither static nor easily quantifiable over time. This has led to the development of potential alternative approaches, such as data trusts, synthetic data, federated learning, and secure multiparty computation [56-59].

Limitations

It should be recognized that there are several notable limitations to our investigation. First, while we relied on state-of-the-art learning architectures, the risk may differ if other computational approaches are considered [21,22]. Second, we did not consider multistage adversarial attacks in which one model is used to predict a demographic, such as sex or age, which is then used to limit the search space, or a scenario in which an adversary manually goes through all potential matches to attempt manual identity verification. However, such approaches would introduce additional uncertainty for the adversary as they would generate predictions for an out-of-sample data set of speech with abnormalities, meaning that accuracy may be lower than expected and the resulting filtered data set may still require many comparisons, in which case our results would apply [60,61]. Third, we did not directly consider the risk of healthy speech versus speech with abnormalities. Nearly all recordings in the Mayo Clinic speech data set contain speech with abnormalities, whereas all VoxCeleb recordings are from healthy speakers. Ideally, there would be a single data set containing both. Fourth, it should be noted that, beyond methodological limitations, our results may not generalize well outside of the United States as the VoxCeleb data have a strong US bias and all the Mayo Clinic recordings were captured in the United States. As such, it will be important to conduct future experiments that leverage alternative computational architectures, more complex adversarial attacks, conversational speech, and data from other geographic regions to assess the

reidentification risk for medical speech data more comprehensively.

In addition, there is an important implication of the VoxCeleb experimental design. As we were interested in a range of set sizes and wanted to complete multiple runs for each size, we combined the train and validation sets from VoxCeleb 1 and 2 and randomly selected a holdout set. However, the ECAPA-TDNN model used for extracting embeddings was pretrained on VoxCeleb, meaning it was exposed to most of the recordings (ie, all but the validation cases) during the original training step [32]. The embeddings are almost certainly superior to what one may have obtained if the embedding model was retrained for each of our splits. Unfortunately, that is not a computationally feasible experimental design. Furthermore, superior embeddings mean we are likely to overestimate risk and draw more conservative conclusions. Given the stakes—reidentification of anonymous research patients—we feel this decision was justified. We also ran a set of experiments using the VoxCeleb validation set as our unknown set (Multimedia Appendix 1). This only allowed for a small unknown set with fixed speakers across runs, so it may be overly optimistic regarding risk. In our opinion, the true risk lies in between our main results and the supplementary results.

Conclusions

In summary, our findings suggest that while the acoustic signal alone can be used for reidentification, the practical risk of reidentification for speech recordings, including elicited recordings typically captured as part of a medical speech examination, is low with sufficiently large search spaces. This risk does vary based on the exact size of the search space—which is dependent on the number of speakers in the known and unknown sets—as well as the similarity of the speech tasks in each set. This provides actionable recommendations to further increase participant privacy and considerations for policy regarding the public release of speech recordings. Finally, we also provide ideas for future studies to extend this work, most notably the need to assess other model architectures and data sets as improvements in speaker identification could substantially increase reidentification risk.

Data Availability

The VoxCeleb 1 and 2 data sets analyzed during this study are available in the VoxCeleb repository [62]. Our Mayo Clinic clinical speech recordings data set analyzed during this study is not publicly available due to the privacy risks related to the release of clinical speech data and are not available by request. We used Python (Python Software Foundation) to implement our code for preprocessing, extracting speaker embeddings, generating subsampled data sets, and running the probabilistic linear discriminant analysis. The source code is available on the internet [63]. The repository also contains detailed documentation for using the scripts.

Acknowledgments

No generative language models were used when writing the manuscript.

Authors' Contributions

DW, BAM, and HB conceived the ideas presented in this study and validated the results. JRD, RLU, and DTJ provided the necessary resources for this study. DW, JLS, and HB developed the methodology for the experiments. DW curated the data for the Mayo Clinic speech recording data set. DW and HB developed the code for running the experiments and visualizing the results. DW conducted formal statistical analysis of the data. DW and HB wrote the original draft of the manuscript. All authors have reviewed and edited the manuscript. DTJ and HB supervised.

Conflicts of Interest

BAM, JRD, RLU, JLS, DTJ, and HB receive funding from the National Institutes of Health. All other authors declare no other conflicts of interest.

Multimedia Appendix 1

An additional set of experiments using only the VoxCeleb validation set as the unknown set. This only allowed for a small unknown set with fixed speakers across runs, so it may be overly optimistic regarding risk. These experiments define a lower bound for risk as compared to the original experiments that draw more conservative conclusions and may overestimate risk.

[[DOCX File, 245 KB - ai_v3i1e52054_app1.docx](#)]

References

1. Fagherazzi G, Fischer A, Ismael M, Despotovic V. Voice for health: the use of vocal biomarkers from research to clinical practice. *Digit Biomark* 2021 Apr 16;5(1):78-88 [FREE Full text] [doi: [10.1159/000515346](https://doi.org/10.1159/000515346)] [Medline: [34056518](https://pubmed.ncbi.nlm.nih.gov/34056518/)]
2. Sara JD, Maor E, Orbelo D, Gulati R, Lerman LO, Lerman A. Noninvasive voice biomarker is associated with incident coronary artery disease events at follow-up. *Mayo Clin Proc* 2022 May;97(5):835-846. [doi: [10.1016/j.mayocp.2021.10.024](https://doi.org/10.1016/j.mayocp.2021.10.024)] [Medline: [35341593](https://pubmed.ncbi.nlm.nih.gov/35341593/)]
3. Maor E, Tsur N, Barkai G, Meister I, Makmel S, Friedman E, et al. Noninvasive vocal biomarker is associated with severe acute respiratory syndrome coronavirus 2 infection. *Mayo Clin Proc Innov Qual Outcomes* 2021 Jun;5(3):654-662 [FREE Full text] [doi: [10.1016/j.mayocpiqo.2021.05.007](https://doi.org/10.1016/j.mayocpiqo.2021.05.007)] [Medline: [34007956](https://pubmed.ncbi.nlm.nih.gov/34007956/)]
4. Tracy JM, Özkanca Y, Atkins DC, Hosseini Ghomi R. Investigating voice as a biomarker: deep phenotyping methods for early detection of Parkinson's disease. *J Biomed Inform* 2020 Apr;104:103362 [FREE Full text] [doi: [10.1016/j.jbi.2019.103362](https://doi.org/10.1016/j.jbi.2019.103362)] [Medline: [31866434](https://pubmed.ncbi.nlm.nih.gov/31866434/)]
5. Asgari M, Shafran I. Predicting severity of Parkinson's disease from speech. *Annu Int Conf IEEE Eng Med Biol Soc* 2010;2010:5201-5204 [FREE Full text] [doi: [10.1109/IEMBS.2010.5626104](https://doi.org/10.1109/IEMBS.2010.5626104)] [Medline: [21095825](https://pubmed.ncbi.nlm.nih.gov/21095825/)]
6. Pignoni A, Delvecchio G, Madonna D, Bressi C, Soares J, Brambilla P. Can Machine Learning help us in dealing with treatment resistant depression? A review. *J Affect Disord* 2019 Dec 01;259:21-26. [doi: [10.1016/j.jad.2019.08.009](https://doi.org/10.1016/j.jad.2019.08.009)] [Medline: [31437696](https://pubmed.ncbi.nlm.nih.gov/31437696/)]
7. GBD 2016 Neurology Collaborators. Global, regional, and national burden of neurological disorders, 1990-2016: a systematic analysis for the Global Burden of Disease Study 2016. *Lancet Neurol* 2019 May;18(5):459-480 [FREE Full text] [doi: [10.1016/S1474-4422\(18\)30499-X](https://doi.org/10.1016/S1474-4422(18)30499-X)] [Medline: [30879893](https://pubmed.ncbi.nlm.nih.gov/30879893/)]
8. GBD 2017 US Neurological Disorders Collaborators, Feigin VL, Vos T, Alahdab F, Amit AM, Bärnighausen TW, et al. Burden of neurological disorders across the US from 1990-2017: a global burden of disease study. *JAMA Neurol* 2021 Feb 01;78(2):165-176 [FREE Full text] [doi: [10.1001/jamaneurol.2020.4152](https://doi.org/10.1001/jamaneurol.2020.4152)] [Medline: [33136137](https://pubmed.ncbi.nlm.nih.gov/33136137/)]
9. Atlas: country resources for neurological disorders 2004: results of a collaborative study of the World Health Organization and the World Federation of Neurology. World Health Organization. 2004. URL: <https://www.who.int/publications/i/item/9241562838> [accessed 2024-02-27]
10. Janca A, Aarli JA, Prilipko L, Dua T, Saxena S, Saraceno B. WHO/WFN Survey of neurological services: a worldwide perspective. *J Neurol Sci* 2006 Aug 15;247(1):29-34. [doi: [10.1016/j.jns.2006.03.003](https://doi.org/10.1016/j.jns.2006.03.003)] [Medline: [16624322](https://pubmed.ncbi.nlm.nih.gov/16624322/)]
11. Duffy JR. *Motor Speech Disorders: Substrates, Differential Diagnosis, and Management*. Amsterdam, The Netherlands: Elsevier Health Sciences; 2019.
12. El Emam K, Jabbouri S, Sams S, Drouet Y, Power M. Evaluating common de-identification heuristics for personal health information. *J Med Internet Res* 2006 Nov 21;8(4):e28 [FREE Full text] [doi: [10.2196/jmir.8.4.e28](https://doi.org/10.2196/jmir.8.4.e28)] [Medline: [17213047](https://pubmed.ncbi.nlm.nih.gov/17213047/)]
13. Ribaric S, Pavešić N. De-identification for privacy protection in biometrics. In: Vielhauer C, editor. *User-Centric Privacy and Security in Biometrics*. London, UK: Institution of Engineering and Technology; 2017:293-324.
14. Malin B, Karp D, Scheuermann RH. Technical and policy approaches to balancing patient privacy and data sharing in clinical and translational research. *J Investig Med* 2023 May 10;58(1):11-18. [doi: [10.2310/jim.0b013e3181c9b2ea](https://doi.org/10.2310/jim.0b013e3181c9b2ea)]
15. Standards for privacy of individually identifiable health information, volume 67. Office for Civil Rights. 2002. URL: <https://www.govinfo.gov/content/pkg/FR-2002-08-14/pdf/02-20554.pdf> [accessed 2024-02-27]
16. Guidance regarding methods for de-identification of protected health information in accordance with the Health Insurance Portability and Accountability Act (HIPAA) privacy rule. Office for Civil Rights. URL: <https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html> [accessed 2024-02-27]

17. Qian J, Han F, Hou J, Zhang C, Wang Y, Li XY. Towards privacy-preserving speech data publishing. In: Proceedings of the 2018 IEEE Conference on Computer Communications. 2018 Presented at: INFOCOM '18; April 16-18, 2018; Honolulu, HI p. 1079-1087 URL: <https://ieeexplore.ieee.org/document/8486250> [doi: [10.1109/infocom.2018.8486250](https://doi.org/10.1109/infocom.2018.8486250)]
18. Atreya RV, Smith JC, McCoy AB, Malin BA, Miller RA. Reducing patient re-identification risk for laboratory results within research datasets. *J Am Med Inform Assoc* 2013 Jan 01;20(1):95-101 [FREE Full text] [doi: [10.1136/amiajnl-2012-001026](https://doi.org/10.1136/amiajnl-2012-001026)] [Medline: [22822040](https://pubmed.ncbi.nlm.nih.gov/22822040/)]
19. Benitez K, Malin B. Evaluating re-identification risks with respect to the HIPAA privacy rule. *J Am Med Inform Assoc* 2010;17(2):169-177 [FREE Full text] [doi: [10.1136/jamia.2009.000026](https://doi.org/10.1136/jamia.2009.000026)] [Medline: [20190059](https://pubmed.ncbi.nlm.nih.gov/20190059/)]
20. Cimino JJ. The false security of blind dates: chrononymization's lack of impact on data privacy of laboratory data. *Appl Clin Inform* 2012;3(4):392-403 [FREE Full text] [doi: [10.4338/ACI-2012-07-RA-0028](https://doi.org/10.4338/ACI-2012-07-RA-0028)] [Medline: [23646086](https://pubmed.ncbi.nlm.nih.gov/23646086/)]
21. Mohd Hanifa R, Isa K, Mohamad S. A review on speaker recognition: technology and challenges. *Comput Electr Eng* 2021 Mar;90:107005. [doi: [10.1016/j.compeleceng.2021.107005](https://doi.org/10.1016/j.compeleceng.2021.107005)]
22. Sztahó D, Szaszák G, Beke A. Deep Learning Methods in Speaker Recognition: A Review. *Period Polytech Electr Eng Comput Sci* 2021 Oct 29;65(4):310-328. [doi: [10.3311/ppee.17024](https://doi.org/10.3311/ppee.17024)]
23. Nagrani A, Chung JS, Xie W, Zisserman A. Voxceleb: large-scale speaker verification in the wild. *Comput Speech Lang* 2020 Mar;60:101027. [doi: [10.1016/j.csl.2019.101027](https://doi.org/10.1016/j.csl.2019.101027)]
24. Lu P, Zhu H, Sovrigno G, Lin X. Voxstructor: voice reconstruction from voiceprint. In: Proceedings of the 24th International Conference on Information Security. 2021 Presented at: ISC '21; November 10-12, 2021; Virtual Event p. 374-397 URL: https://link.springer.com/chapter/10.1007/978-3-030-91356-4_20 [doi: [10.1007/978-3-030-91356-4_20](https://doi.org/10.1007/978-3-030-91356-4_20)]
25. Qian J, Du H, Hou J, Chen L, Jung T, Li XY. Speech sanitizer: speech content desensitization and voice anonymization. *IEEE Trans Dependable Secure Comput* 2021 Nov 1;18(6):2631-2642. [doi: [10.1109/tdsc.2019.2960239](https://doi.org/10.1109/tdsc.2019.2960239)]
26. Arasteh ST, Weise T, Schuster M, Noeth E, Maier A, Yang SH. The effect of speech pathology on automatic speaker verification: a large-scale study. *Sci Rep* 2022;13:20476. [doi: [10.1038/s41598-023-47711-7](https://doi.org/10.1038/s41598-023-47711-7)]
27. Young V, Mihailidis A. Difficulties in automatic speech recognition of dysarthric speakers and implications for speech-based applications used by the elderly: a literature review. *Assist Technol* 2010 Jun 10;22(2):99-114. [doi: [10.1080/10400435.2010.483646](https://doi.org/10.1080/10400435.2010.483646)] [Medline: [20698428](https://pubmed.ncbi.nlm.nih.gov/20698428/)]
28. Mustafa MB, Rosdi F, Salim SS, Mughal MU. Exploring the influence of general and specific factors on the recognition accuracy of an ASR system for dysarthric speaker. *Expert Syst Appl* 2015 May;42(8):3924-3932. [doi: [10.1016/j.eswa.2015.01.033](https://doi.org/10.1016/j.eswa.2015.01.033)]
29. De Russis L, Corno F. On the impact of dysarthric speech on contemporary ASR cloud platforms. *J Reliable Intell Environ* 2019 Jul 6;5(3):163-172. [doi: [10.1007/s40860-019-00085-y](https://doi.org/10.1007/s40860-019-00085-y)]
30. Lévêque N, Slis A, Lancia L, Bruneteau G, Fougeron C. Acoustic change over time in spastic and/or flaccid dysarthria in motor neuron diseases. *J Speech Lang Hear Res* 2022 May 11;65(5):1767-1783. [doi: [10.1044/2022.jslhr-21-00434](https://doi.org/10.1044/2022.jslhr-21-00434)]
31. Dankar FK, El Emam KE. A method for evaluating marketer re-identification risk. In: Proceedings of the 2010 EDBT/ICDT Workshops. 2010 Presented at: EDBT '10; March 22-26, 2010; Lausanne, Switzerland p. 1-10 URL: <https://dl.acm.org/doi/10.1145/1754239.1754271> [doi: [10.1145/1754239.1754271](https://doi.org/10.1145/1754239.1754271)]
32. Desplanques B, Thienpondt J, Demuyneck K. ECAPA-TDNN: emphasized channel attention, propagation and aggregation in TDNN based speaker verification. In: Proceedings of the 2020 Technical Conference Focused on Speech Processing and Application. 2020 Presented at: INTERSPEECH '20; October 25-29, 2020; Virtual Event p. 3830-3834 URL: https://www.isca-archive.org/interspeech_2020/desplanques20_interspeech.html [doi: [10.21437/interspeech.2020-2650](https://doi.org/10.21437/interspeech.2020-2650)]
33. Khosravani A, Homayounpour MM. A PLDA approach for language and text independent speaker recognition. *Comput Speech Lang* 2017 Sep;45:457-474. [doi: [10.1016/j.csl.2017.04.003](https://doi.org/10.1016/j.csl.2017.04.003)]
34. Borgström BJ. Discriminative training of PLDA for speaker verification with x-vectors. Department of Defense Under Air Force. URL: <https://www.ll.mit.edu/sites/default/files/publication/doc/discriminative-PLDA-speaker-verification-borgstrom-121037.pdf> [accessed 2024-02-27]
35. Nagrani A, Chung JS, Zisserman A. VoxCeleb: a large-scale speaker identification dataset. In: Proceedings of the 18th Annual Conference of the International Speech Communication Association. 2017 Presented at: Interspeech '17; August 20-24, 2017; Stockholm, Sweden p. 2616-2620 URL: https://www.isca-archive.org/interspeech_2017/nagrani17_interspeech.html [doi: [10.21437/interspeech.2017-950](https://doi.org/10.21437/interspeech.2017-950)]
36. Chung JS, Nagrani A, Zisserman A. VoxCeleb2: deep speaker recognition. In: Proceedings of the 19th Annual Conference of the International Speech Communication Association. 2018 Presented at: INTERSPEECH '18; September 2-6, 2018; Hyderabad, India p. 1086-1090 URL: https://www.isca-archive.org/interspeech_2018/chung18b_interspeech.html [doi: [10.21437/interspeech.2018-1929](https://doi.org/10.21437/interspeech.2018-1929)]
37. Ibrahim NS, Ramli DA. I-vector extraction for speaker recognition based on dimensionality reduction. *Procedia Comput Sci* 2018;126:1534-1540. [doi: [10.1016/j.procs.2018.08.126](https://doi.org/10.1016/j.procs.2018.08.126)]
38. Wilkinghoff K. On open-set speaker identification with I-vectors. In: Proceedings of the 2020 conference on Speaker and Language Recognition Workshop. 2020 Presented at: Odyssey' 20; November 2-5, 2020; Tokyo, Japan p. 408-414 URL: https://www.isca-archive.org/odyssey_2020/wilkinghoff20_odyssey.html [doi: [10.21437/odyssey.2020-58](https://doi.org/10.21437/odyssey.2020-58)]

39. Ravanelli M, Parcollet T, Plantinga P, Rouhe A, Cornell S, Lugosch L, et al. SpeechBrain: a general-purpose speech toolkit. arXiv Preprint posted online June 8, 2021 [[FREE Full text](#)]
40. Prince SJ, Elder JH. Probabilistic linear discriminant analysis for inferences about identity. In: Proceedings of the 2007 IEEE 11th International Conference on Computer Vision. 2007 Presented at: ICCV '07; October 14-21, 2007; Rio de Janeiro, Brazil p. 1-8 URL: <https://ieeexplore.ieee.org/document/4409052> [doi: [10.1109/iccv.2007.4409052](https://doi.org/10.1109/iccv.2007.4409052)]
41. Mahajan D, Ramamoorthi R, Curless B. A theory of spherical harmonic identities for BRDF/lighting transfer and image consistency. In: Proceedings of the 9th European Conference on Computer Vision on Computer Vision. 2006 Presented at: ECCV '06; May 7-13, 2006; Graz, Austria p. 41-55 URL: https://link.springer.com/chapter/10.1007/11744085_4 [doi: [10.1007/11744085_4](https://doi.org/10.1007/11744085_4)]
42. van Leeuwen DA, Brümmer N. An introduction to application-independent evaluation of speaker recognition systems. In: Müller C, editor. Speaker Classification I: Fundamentals, Features, and Methods. Berlin, Germany: Springer; 2007.
43. Van Noorden R. The ethical questions that haunt facial-recognition research. *Nature* 2020 Nov 18;587(7834):354-358. [doi: [10.1038/d41586-020-03187-3](https://doi.org/10.1038/d41586-020-03187-3)] [Medline: [33208967](https://pubmed.ncbi.nlm.nih.gov/33208967/)]
44. Social media fact sheet. Pew Research Center. 2021. URL: <https://www.pewresearch.org/internet/fact-sheet/social-media/?menuItem=4abfc543-4bd1-4b1f-bd4a-e7c67728ab76> [accessed 2024-02-27]
45. Zhang L, Duvvuri R, Chandra KK, Nguyen T, Ghomi RH. Automated voice biomarkers for depression symptoms using an online cross-sectional data collection initiative. *Depress Anxiety* 2020 Jul 07;37(7):657-669. [doi: [10.1002/da.23020](https://doi.org/10.1002/da.23020)] [Medline: [32383335](https://pubmed.ncbi.nlm.nih.gov/32383335/)]
46. Sharma N, Krishnan P, Kumar R, Ramoji S, Chetupalli SR, Ghosh PK, et al. Coswara — a database of breathing, cough, and voice sounds for COVID-19 diagnosis. In: Proceedings of the 2020 Technical Conference Focused on Speech Processing and Application. 2020 Presented at: INTERSPEECH '20; October 25-29, 2020; Virtual Event p. 4811-4815 URL: https://www.isca-archive.org/interspeech_2020/sharma20d_interspeech.html [doi: [10.21437/interspeech.2020-2768](https://doi.org/10.21437/interspeech.2020-2768)]
47. Turrisi R, Braccia A, Emanuele M, Giulietti S, Pugliatti M, Sensi M, et al. EasyCall corpus: a dysarthric speech dataset. In: Proceedings of the 2021 Technical Conference Focused on Speech Processing and Application. 2021 Presented at: INTERSPEECH '21; August 30-September 3, 2021; Brno, Czech Republic p. 41-45 URL: https://www.isca-archive.org/interspeech_2021/turrisi21_interspeech.html [doi: [10.21437/interspeech.2021-549](https://doi.org/10.21437/interspeech.2021-549)]
48. Flechl M, Yin SC, Park J, Skala P. End-to-end speech recognition modeling from de-identified data. In: Proceedings of the 2022 Technical Conference Focused on Speech Processing and Application. 2022 Presented at: INTERSPEECH '22; September 18-22, 2022; Incheon, South Korea p. 1382-1386 URL: https://www.isca-archive.org/interspeech_2022/flechl22_interspeech.html [doi: [10.21437/interspeech.2022-10484](https://doi.org/10.21437/interspeech.2022-10484)]
49. Kenny P, Boulianne G, Ouellet P, Dumouchel P. Speaker and session variability in GMM-based speaker verification. *IEEE Trans Audio Speech Lang Process* 2007 May;15(4):1448-1460. [doi: [10.1109/tasl.2007.894527](https://doi.org/10.1109/tasl.2007.894527)]
50. Kinnunen T, Li H. An overview of text-independent speaker recognition: from features to supervectors. *Speech Commun* 2010 Jan;52(1):12-40. [doi: [10.1016/j.specom.2009.08.009](https://doi.org/10.1016/j.specom.2009.08.009)]
51. Liu Y, Yan C, Yin Z, Wan Z, Xia W, Kantarcioglu M, et al. Biomedical research cohort membership disclosure on social media. *AMIA Annu Symp Proc* 2019;2019:607-616 [[FREE Full text](#)] [Medline: [32308855](https://pubmed.ncbi.nlm.nih.gov/32308855/)]
52. Xia W, Liu Y, Wan Z, Vorobeychik Y, Kantarcioglu M, Nyemba S, et al. Enabling realistic health data re-identification risk assessment through adversarial modeling. *J Am Med Inform Assoc* 2021 Mar 18;28(4):744-752 [[FREE Full text](#)] [doi: [10.1093/jamia/ocaa327](https://doi.org/10.1093/jamia/ocaa327)] [Medline: [33448306](https://pubmed.ncbi.nlm.nih.gov/33448306/)]
53. Tucker K, Branson J, Dilleen M, Hollis S, Loughlin P, Nixon MJ, et al. Protecting patient privacy when sharing patient-level data from clinical trials. *BMC Med Res Methodol* 2016 Jul 08;16 Suppl 1(Suppl 1):77 [[FREE Full text](#)] [doi: [10.1186/s12874-016-0169-4](https://doi.org/10.1186/s12874-016-0169-4)] [Medline: [27410040](https://pubmed.ncbi.nlm.nih.gov/27410040/)]
54. Culnane B, Rubinstein BI, Teague V. Health data in an open world. arXiv Preprint posted online December 15, 2017 2017. [doi: [10.48550/arXiv.1712.05627](https://doi.org/10.48550/arXiv.1712.05627)]
55. Rocher L, Hendrickx JM, de Montjoye Y. Estimating the success of re-identifications in incomplete datasets using generative models. *Nat Commun* 2019 Jul 23;10(1):3069 [[FREE Full text](#)] [doi: [10.1038/s41467-019-10933-3](https://doi.org/10.1038/s41467-019-10933-3)] [Medline: [31337762](https://pubmed.ncbi.nlm.nih.gov/31337762/)]
56. Young M, Rodriguez L, Keller E, Sun F, Sa B, Whittington J, et al. Beyond open vs. closed: balancing individual privacy and public accountability in data sharing. In: Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency. 2019 Presented at: FAT* '19; January 29-31, 2019; Atlanta, GA p. 191-200 URL: <https://dl.acm.org/doi/abs/10.1145/3287560.3287577> [doi: [10.1145/3287560.3287577](https://doi.org/10.1145/3287560.3287577)]
57. Rieke N, Hancox J, Li W, Milletari F, Roth HR, Albarqouni S, et al. The future of digital health with federated learning. *NPJ Digit Med* 2020 Sep 14;3(1):119 [[FREE Full text](#)] [doi: [10.1038/s41746-020-00323-1](https://doi.org/10.1038/s41746-020-00323-1)] [Medline: [33015372](https://pubmed.ncbi.nlm.nih.gov/33015372/)]
58. Kaissis GA, Makowski MR, Rückert D, Braren RF. Secure, privacy-preserving and federated machine learning in medical imaging. *Nat Mach Intell* 2020 Jun 08;2(6):305-311. [doi: [10.1038/s42256-020-0186-1](https://doi.org/10.1038/s42256-020-0186-1)]
59. Ng D, Lan X, Yao MM, Chan WP, Feng M. Federated learning: a collaborative effort to achieve better medical imaging models for individual sites that have small labelled datasets. *Quant Imaging Med Surg* 2021 Feb;11(2):852-857 [[FREE Full text](#)] [doi: [10.21037/qims-20-595](https://doi.org/10.21037/qims-20-595)] [Medline: [33532283](https://pubmed.ncbi.nlm.nih.gov/33532283/)]
60. Xia W, Kantarcioglu M, Wan Z, Heatherly RD, Vorobeychik Y, Malin BA. Process-driven data privacy. In: Proceedings of the 24th ACM International on Conference on Information and Knowledge Management. 2015 Presented at: CIKM '15;

- October 18-23, 2015; Melbourne, Australia p. 1021-1030 URL: <https://dl.acm.org/doi/10.1145/2806416.2806580> [doi: [10.1145/2806416.2806580](https://doi.org/10.1145/2806416.2806580)]
61. Venkatesaramani R, Malin BA, Vorobeychik Y. Re-identification of individuals in genomic datasets using public face images. *Sci Adv* 2021 Nov 19;7(47):eabg3296 [FREE Full text] [doi: [10.1126/sciadv.abg3296](https://doi.org/10.1126/sciadv.abg3296)] [Medline: [34788101](https://pubmed.ncbi.nlm.nih.gov/34788101/)]
 62. VoxCeleb: a large scale audio-visual dataset of human speech. Visual Geometry Group. URL: <https://www.robots.ox.ac.uk/~vgg/data/voxceleb/> [accessed 2024-03-05]
 63. Wiepert DA, Malin BA, Duffy JR, Utianski RL, Stricker JL, Jones DT, et al. Risk of re-identification for shared clinical speech recordings. GitHub. URL: https://github.com/Neurology-AI-Program/Speech_risk [accessed 2024-03-05]

Abbreviations

AMR: alternating motion rate

ECAPA-TDNN: Emphasized Channel Attention, Propagation, and Aggregation in Time-Delay Neural Network

EER: equal error rate

FA: false acceptance

FAR: false acceptance rate

FR: false rejection

HIPAA: Health Insurance Portability and Accountability Act

minDCF: minimum detection cost function

PLDA: probabilistic linear discriminant analysis

SMR: sequential motion rate

TA: true acceptance

Edited by A Mavragani; submitted 21.08.23; peer-reviewed by M Abdalla, Y Khan, E Toki; comments to author 07.12.23; revised version received 26.01.24; accepted 19.02.24; published 15.03.24.

Please cite as:

*Wiepert D, Malin BA, Duffy JR, Utianski RL, Stricker JL, Jones DT, Botha H
Reidentification of Participants in Shared Clinical Data Sets: Experimental Study
JMIR AI 2024;3:e52054*

URL: <https://ai.jmir.org/2024/1/e52054>

doi: [10.2196/52054](https://doi.org/10.2196/52054)

PMID: [38875581](https://pubmed.ncbi.nlm.nih.gov/38875581/)

©Daniela Wiepert, Bradley A Malin, Joseph R Duffy, Rene L Utianski, John L Stricker, David T Jones, Hugo Botha. Originally published in JMIR AI (<https://ai.jmir.org/>), 15.03.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR AI, is properly cited. The complete bibliographic information, a link to the original publication on <https://www.ai.jmir.org/>, as well as this copyright and license information must be included.

Original Paper

Predictive Modeling of Hypertension-Related Postpartum Readmission: Retrospective Cohort Analysis

Jinxin Tao¹, PhD; Ramsey G Larson², MD; Yonatan Mintz¹, PhD; Oguzhan Alagoz¹, PhD; Kara K Hoppe³, MS, DO

¹Industrial and Systems Engineering, University of Wisconsin Madison, Madison, WI, United States

²Department of Obstetrics and Gynecology, MultiCare Rockwood Clinic, Spokane, WA, United States

³Department of Obstetrics and Gynecology, School of Medicine and Public Health, University of Wisconsin Madison, Madison, WI, United States

Corresponding Author:

Yonatan Mintz, PhD

Industrial and Systems Engineering

University of Wisconsin Madison

3107 Mechanical Engineering Building

1513 University Avenue

Madison, WI, 53706

United States

Phone: 1 6092180891

Email: ymintz@wisc.edu

Abstract

Background: Hypertension is the most common reason for postpartum hospital readmission. Better prediction of postpartum readmission will improve the health care of patients. These models will allow better use of resources and decrease health care costs.

Objective: This study aimed to evaluate clinical predictors of postpartum readmission for hypertension using a novel machine learning (ML) model that can effectively predict readmissions and balance treatment costs. We examined whether blood pressure and other measures during labor, not just postpartum measures, would be important predictors of readmission.

Methods: We conducted a retrospective cohort study from the PeriData website data set from a single midwestern academic center of all women who delivered from 2009 to 2018. This study consists of 2 data sets; 1 spanning the years 2009-2015 and the other spanning the years 2016-2018. A total of 47 clinical and demographic variables were collected including blood pressure measurements during labor and post partum, laboratory values, and medication administration. Hospital readmissions were verified by patient chart review. In total, 32,645 were considered in the study. For our analysis, we trained several cost-sensitive ML models to predict the primary outcome of hypertension-related postpartum readmission within 42 days post partum. Models were evaluated using cross-validation and on independent data sets (models trained on data from 2009 to 2015 were validated on the data from 2016 to 2018). To assess clinical viability, a cost analysis of the models was performed to see how their recommendations could affect treatment costs.

Results: Of the 32,645 patients included in the study, 170 were readmitted due to a hypertension-related diagnosis. A cost-sensitive random forest method was found to be the most effective with a balanced accuracy of 76.61% for predicting readmission. Using a feature importance and area under the curve analysis, the most important variables for predicting readmission were blood pressures in labor and 24-48 hours post partum increasing the area under the curve of the model from 0.69 (SD 0.06) to 0.81 (SD 0.06), ($P=.05$). Cost analysis showed that the resulting model could have reduced associated readmission costs by US \$6000 against comparable models with similar F_1 -score and balanced accuracy. The most effective model was then implemented as a risk calculator that is publicly available. The code for this calculator and the model is also publicly available at a GitHub repository.

Conclusions: Blood pressure measurements during labor through 48 hours post partum can be combined with other variables to predict women at risk for postpartum readmission. Using ML techniques in conjunction with these data have the potential to improve health outcomes and reduce associated costs. The use of the calculator can greatly assist clinicians in providing care to patients and improve medical decision-making.

(JMIR AI 2024;3:e48588) doi:[10.2196/48588](https://doi.org/10.2196/48588)

KEYWORDS

pregnancy; postpartum; hypertension; preeclampsia; blood pressure; hospital readmission, clinical calculator; healthcare cost; cost; cohort analysis; utilization; resources; labor; women; risk; readmission; cohort; hospital; statistical model; retrospective cohort study; predict; risk

Introduction

Hypertensive disorders of pregnancy (HDP) are common and estimated to occur in 10% of pregnancies in the United States. In addition to complicating the management of pre- and postdelivery periods, hypertension is the leading cause of postpartum readmission, accounting for 9.3%-27% of postpartum readmissions [1-3]. Postpartum readmission is costly, both in health care dollars and in quality-of-life measures for mothers and new families. In addition, HDP increases maternal morbidity and mortality and is associated with an increased risk of cardiovascular disease later in life [4-8].

A study published in the *Journal of Hypertension* in 2018 estimated that preventable postpartum readmission in women with hypertension resulted in 20,240 excess inpatient hospital days and US \$36 million excess medical costs [9]. Rates and reasons for readmission have been under recent scrutiny and offer an area to improve health care delivery and preventative care. All-cause hospital readmission rates are on the rise with risk factors for all-cause postpartum readmission including public insurance, race, presence of comorbid conditions including hypertension and diabetes, and cesarean section [3].

Approximately 30% of women who experience hypertension-related postpartum readmissions do not have antecedent diagnoses of hypertension, thus making it imperative to include normotensive patients without an HDP before postpartum discharge in evaluating for postpartum readmission [10]. As such, our objective was to identify key clinical variables, in addition to demographic characteristics, implicated in postpartum readmission of all birthing persons using a machine learning (ML) model. This prediction task is challenging because while costs related to readmission are high, readmission rates are low resulting in highly imbalanced data sets that are challenging to use in training ML models. For instance, while existing models have strong overall accuracy performance (out of sample, an area under the curve of 0.81) [11], they do so by trading off high specificity for low sensitivity which could result in many readmission cases going undetected and not properly treated. We hypothesized that blood pressure metrics during labor, not just post partum, would impact readmission rates. Similarly, we hypothesized that antihypertensive medication administration and high preeclampsia laboratory values during initial readmission would increase the readmission rate.

Methods

Ethical Considerations

We obtained institutional review board approval (#2016-006). Individual patient consent was not required due to the retrospective study design. The data set was deidentified before study analysis. No compensation was provided to human

participants as this was a retrospective study that involved development of a retrospective data set using electronic medical records (EMRs).

Chart Review and Inclusion

We initially performed a retrospective chart review of all patients who delivered at a single, midwestern academic center hospital between 2009 and 2015. Inclusion criteria for this study included all women who delivered a baby in this time frame. We wanted to ensure that we captured all hypertension-related readmissions within 42 days post partum regardless of a diagnosis of hypertension before hospital discharge from the delivery of their infant. The primary outcome was hypertension-related readmission; therefore, all readmissions included in this data set were specific to hypertension only. To confirm our previous results and create a larger sample size, we extended the study population to include births from 2016 to 2018 and used a similar process to manage the data of all patients who delivered at the same birthing hospital. We used the hospital's PeriData website [12] data set, which is used to contribute birth-related outcomes to the state-wide database for clinical perinatal information and additional hospital-run reports to obtain additional data available from the EMR. We collected demographic as well as clinical data, including blood pressure measurements during labor and post partum, laboratory data, and medication administration at the patient level to be our predictor variables. Hospital readmissions and our prediction response variable were verified by patient chart review. Given that the data came from multiple sources and had missing observations, the raw data set could not be used directly for analysis.

Analytics Plan

Data Processing and Feature Engineering

We processed the raw data set and then merged the processed data including patient demographics, blood pressure measurements, medication administration, and laboratory information from different sources into 1 pandas data frame [13]. Race and ethnicity were entered in the medical chart based on the patient's self-identity at the time of admission to the health care system. Laboratory results were included in this analysis because they are involved in the classification and severity of HDP. Laboratory results included liver function tests, hemoglobin and platelet counts, creatinine, and urine protein. We analyzed blood pressure records with timestamps and identified the highest systolic blood pressure and associated diastolic blood pressure during 3 time periods, that are, labor, 0-24 hours post partum, and 24-48 hours post partum, because we expected blood pressure during labor and post partum to be important features for predicting hypertensive readmissions. Using the medication administration data from the EMR, we constructed the following binary (yes or no) attributes for the following medication name and route administered: (1) oral labetalol, (2) intravenous labetalol, (3) oral nifedipine-immediate

release, (4) oral nifedipine-extended release, (5) intravenous hydralazine, and (6) oral ibuprofen. To obtain these features, we started with a full medication data set for each patient's medical registration number that included the medication name, time administered, dosage, and route of administration. This meant that there were multiple entries per medical registration number if a particular patient was given that medication more than once. The key challenge of using the medication data was that there were significant missing data; moreover, not all patients received all medications, and the individual medication schedules could be infrequent. For this reason, we considered only the binary attributes instead of the full medication schedule to ensure that the data points were dense enough for the analysis.

Predictive Model Training and Validation

We used a cost-sensitive random forest method to predict which patient would experience a hypertension-related postpartum readmission [14]. Since the data set was imbalanced (only 170 readmissions out of 32,645 participants), the use of class weights that penalize false negatives significantly higher than false positives was necessary to avoid ML models that predict every sample as the negative class. We considered other candidate classifiers namely logistic regression with L1 or L2 regularization, support vector machines (SVM) with polynomial, radial basis function or sigmoid kernel, and a standard decision tree approach for the prediction task. To measure the predictive performance of each model, we considered a combination of different metrics. In the case of imbalanced data, reporting high accuracy may be inappropriate since a highly accurate model could simply ignore the rare class and still achieve high accuracy. Therefore, we considered 2 complementary scores for assessing our model namely balanced accuracy and the F_1 -score [15]. The balanced accuracy can be thought of as balancing the frequency of true positives and true negatives. When calculating accuracy, it can be calculated by averaging the true negative rate (specificity) and the true positive rate (sensitivity) of the model. In addition to prediction accuracy, since our setting has a low frequency of positive cases, we needed to ensure our selected model had high precision (alternatively low false alarm rate). For that reason, we also considered the F_1 -score, which measures the balance between the precision and the true positive rate. We tuned the hyperparameters of each model using cross-validation. For the outer loop, we iterated every hyperparameter combination. Then we performed stratified 5-fold cross-validation in the inner loop and optimized the hyperparameters by evaluating the average balanced accuracy. Each model was trained using its respective classifier implementation from scikit-learn [16].

We trained models on the 2009-2015 data and 2016-2018 data individually and validated them using the 5-fold cross-validation

pipeline. The purpose of this was to see if different factors impacted readmission rates and decisions between the 2 time periods. For added validation, we computed the performance of models trained on the 2009-2015 data set using the 2016-2018 data to evaluate our pipeline. The final model deployed in practice was tuned using the combined data set and 5-fold cross-validation. We performed a feature importance analysis on the best models chosen by cross-validation for each data set.

Cost Analysis and Estimating Clinical Impact

To estimate the clinical impact of predictions, we completed 2 different forms of cost analysis. For each candidate model considered, we used the above cross-validation procedure to compute their estimated implementation costs.

We estimated the value of a false negative (an unplanned or unpredicted readmission) to be US \$20,439 and the value of a false positive (the price of labetalol for 6 weeks for a patient who ultimately did not need it) to be US \$36. These costs were based on estimates derived from our previous research [17]. In addition, for the cost-sensitive random forest model (which we ultimately determined was the most effective model), we performed an additional analysis. For this analysis, we took the model's score for how likely a patient was to be classified as needing readmission and compared it with a predictive threshold. If the model score was larger than the threshold, the model would predict that the patient would be readmitted. When the threshold is <0.5 , more patients are predicted to be readmitted and if the threshold is >0.5 , more patients are predicted to not be readmitted. We used leave-one-out cross-validation to compute the overall medical costs and balanced accuracy for different thresholds between 0 and 1. The goal of this analysis was to see how model scores should be interpreted in practice by decision makers so that overall medical costs are minimized.

Results

Data Overview

From January 2009 to December 2018, a total of 39,133 women delivered at our hospital; however, only 32,645 had complete medical records available for analysis. Of these, 170 women were readmitted for a hypertension-related diagnosis. There was a statistically significant difference between the readmitted group and the not readmitted group in terms of maternal age, gestational age at delivery, race, BMI, mode of delivery, and hypertension diagnosis. The readmitted group was more likely to be older, having earlier gestational age at delivery, Black race, higher BMI, cesarean delivery, and having a diagnosis of chronic or pregnancy-induced hypertension (Table 1). The rate of hypertension diagnosis in our sample was 9%. The rate of readmission was 0.5%.

Table 1. Patient demographics and comparisons between the readmitted group and the not readmitted group.

Characteristics	All patients (N=32,645)	Readmitted (n=170)	Not readmitted (n=32,475)	P value
Maternal age, mean (SD)	30.5 (5.3)	32.9 (5.7)	30.5 (5.3)	<.001
Parity, n (%)				.29
Nulliparous	10,786 (33)	61 (35.9)	10,725 (33)	
Multiparous	17,946 (55)	95 (55.9)	17,851 (55)	
Unknown	3913 (12)	14 (8.2)	3899 (12)	
Gestational age at delivery in weeks, mean (SD)	38.9 (2.4)	37.7 (2.5)	38.9 (2.4)	<.001
Race, n (%)				<.001
White	26,188 (80.2)	130 (76.5)	26,058 (80.2)	
Black	1221 (3.7)	18 (10.6)	1203 (3.7)	
Asian Indian	2532 (7.8)	12 (7.1)	2520 (7.8)	
Asian, other	855 (2.6)	3 (1.8)	852 (2.6)	
American Indian or Native	457 (1.4)	1 (0.6)	456 (1.4)	
Native Hawaiian	67 (0.2)	0 (0)	67 (0.2)	
Unknown or other	1325 (4.1)	6 (3.5)	1319 (4.1)	
Hispanic, n (%)				.05
Yes	2937 (9)	8 (4.7)	2929 (9)	
No	29,708 (91)	162 (95.3)	29,546 (91)	
BMI ^a , mean (SD)	26.5 (8.6)	28.7 (8.6)	26.5 (8.9)	.001
Mode of delivery, n (%)				<.001
Vaginal	20,217 (61.9)	76 (44.7)	20,141 (62)	
Vaginal vacuum	1695 (5.2)	8 (4.7)	1687 (5.2)	
Vaginal forceps	582 (1.8)	3 (1.8)	579 (1.8)	
Cesarean section	10,151 (31.1)	83 (48.8)	10,068 (31)	
Hypertension diagnosis, n (%)	2952 (9)	96 (56.5)	2856 (8.8)	<.001
Chronic hypertension				
Without preeclampsia	299 (1)	16 (9.4)	283 (0.9)	
With preeclampsia	284 (0.9)	15 (8.8)	269 (0.8)	
Gestational hypertension	793 (2.4)	13 (7.6)	780 (2.4)	
Preeclampsia				
Mild	582 (1.8)	26 (15.3)	556 (1.7)	
Severe	828 (2.5)	18 (10.6)	810 (2.5)	
Unspecified	166 (0.5)	8 (4.7)	158 (0.5)	

^aBMI: weight in kilograms divided by the square of height in meters.

Predictive Model Results

During our initial analysis of the data from 2009 to 2015, we evaluated 47 clinical and demographic variables to assess their importance in predicting postpartum readmission (Figure 1). The variables most important for predicting readmission included blood pressure parameters during labor and through the postpartum period as well as factors such as prepregnancy BMI, maternal age, and gestational age at delivery. Variables that had less predictive value included an HDP, administration

of antihypertensive medication, and mode of delivery. To increase the predictive accuracy of the model, many of these variables were excluded from the next analysis. Even with fewer variables, again the diagnosis of HDP and mode of delivery were of least importance and blood pressure data during labor and post partum were most important. Additional details on the model feature importance and feature correlations can be found in the Multimedia Appendix 1 [18]. Through cross-validation analysis, we found that the best model in terms of balanced accuracy and F_1 -score was the random forest model. We

performed an additional validation by using the best-tuned models from the 2009 to 2015 data set on the 2016 to 2018 data sets; the results are shown in Table 2. Each model was trained only on the 2009-2015 data and was used to predict readmission for patients in the 2016-2018 data. As shown in Table 2, the

random forest model achieves the best-balanced accuracy and F_1 -scores among all candidate models. Please refer to the Multimedia Appendix 1 [18] for the full set of cross-validation parameters for each model.

Figure 1. Feature importance plot for hypertension-related postpartum readmission, 2009-2015. APGAR: appearance, pulse, grimace response, activity, respiration; DBP: diastolic blood pressure; NICU: neonatal intensive care unit; SBP: systolic blood pressure.

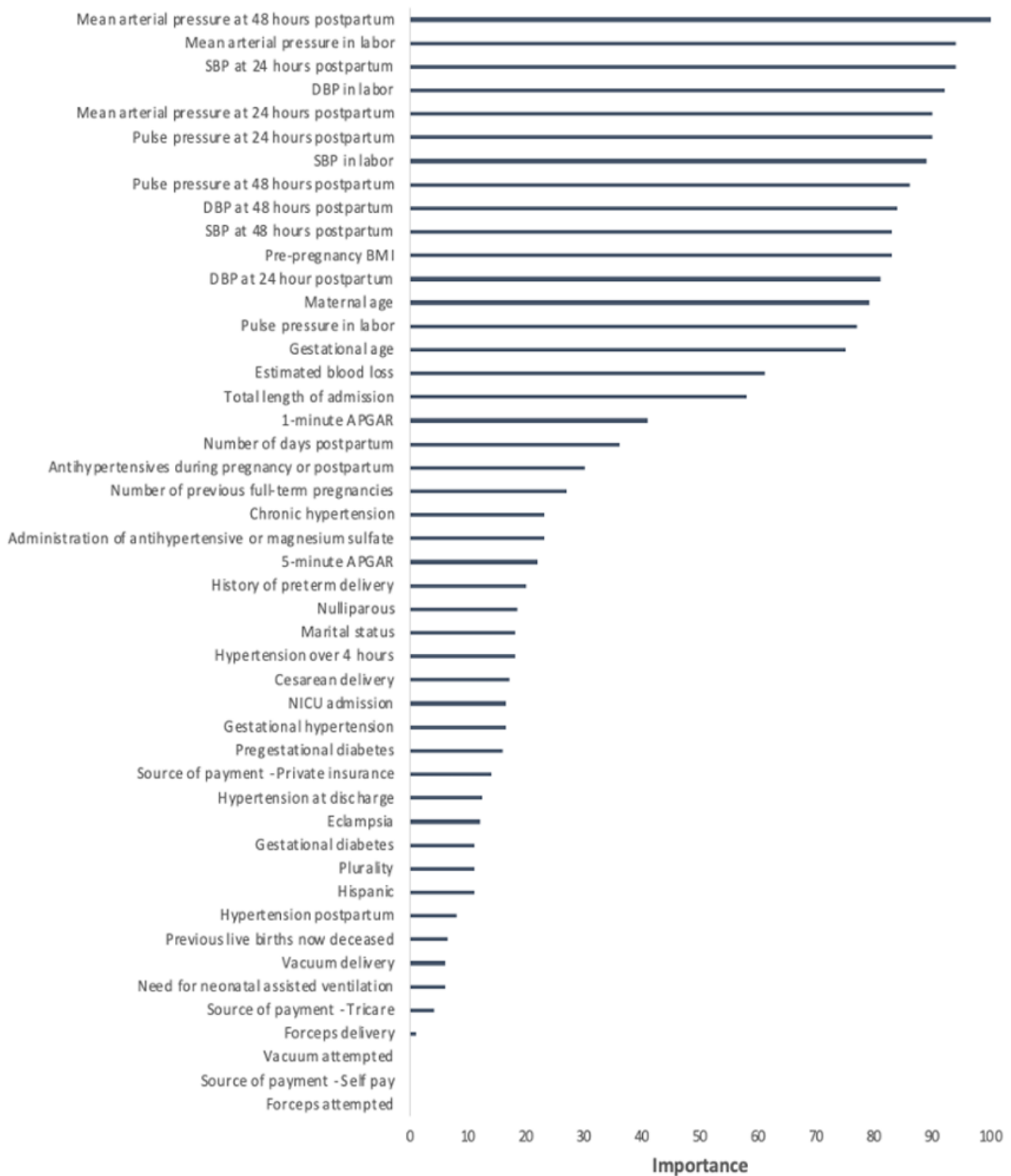


Table 2. Predictive performance of models trained using the 2009-2015 data set on the 2016 to 2018 data sets.

Model type	Specificity, %	Sensitivity, %	Precision or PPV ^a , %	NPV ^b , %	F ₁ -score	Balanced accuracy, %	Cost (US \$)
Random forest	70.86	75.81	1.38	99.82	0.027	73.33	426,240
Decision tree	64.94	75.81	1.15	99.8	0.023	70.37	450,828
SVM ^c	53.21	90.32	1.03	99.9	0.020	71.77	316,512
Logistic regression L1	62.12	83.87	1.17	99.86	0.023	72.99	360,864
Logistic regression L2	63.09	83.87	1.21	99.86	0.024	73.48	356,832

^aPPV: positive predictive value.

^bNPV: negative predictive value.

^cSVM: support vector machines.

Feature Importance Analysis

The additional 11,608 participants from deliveries between 2016 and 2018 were then added to the data set, and medication administration data and laboratory data were included in the next analysis. The final data set included 32,645 patients. Out of 33,482 total patients, 837 were excluded from the analysis because of incomplete information regarding key features. We ranked the features by their predictive importance and selected the final set of features to be (1) BMI; (2) gestational age at delivery; (3) maternal age; (4) highest systolic blood pressure during 3 time periods, that were labor, 0-24 hours post partum, and 24-48 hours post partum; and (5) binary medication features. The laboratory features were discarded because of their low predictive feature importance. The most important clinical variable in predicting readmission was systolic blood pressure

between 24 and 48 hours post partum, and the second most important was systolic blood pressure during labor (Figure 2). Other factors that continued to be of importance in predicting readmission included gestational age at delivery, maternal age, and prepregnancy BMI. We computed the correlation between our proposed features (Figure 3). The receiver operating characteristic (ROC) curves demonstrate that our model is able to distinguish between a true positive (meaning predicting a readmission) and a false positive (meaning incorrectly predicting a readmission). To show the significance of blood pressure features in readmission prediction, we did a ROC curve comparison using a 10-fold cross-validation with and without blood pressure features and calculated the mean ROC and associated SD, respectively. We can see a significant decline in the classification performance without blood pressure features (Figure 4).

Figure 2. Validated feature importance plot for hypertension-related postpartum readmission, 2009-2018 (blood pressure values were the highest recorded values during the specified time frames).

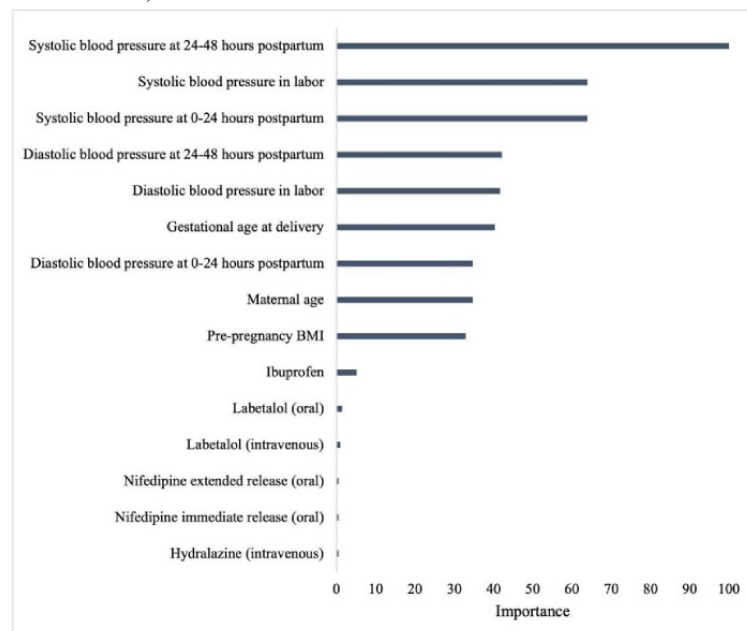


Figure 3. Correlation matrix between features on the combined data from 2009 to 2018. IV: intravenous; PO: orally.

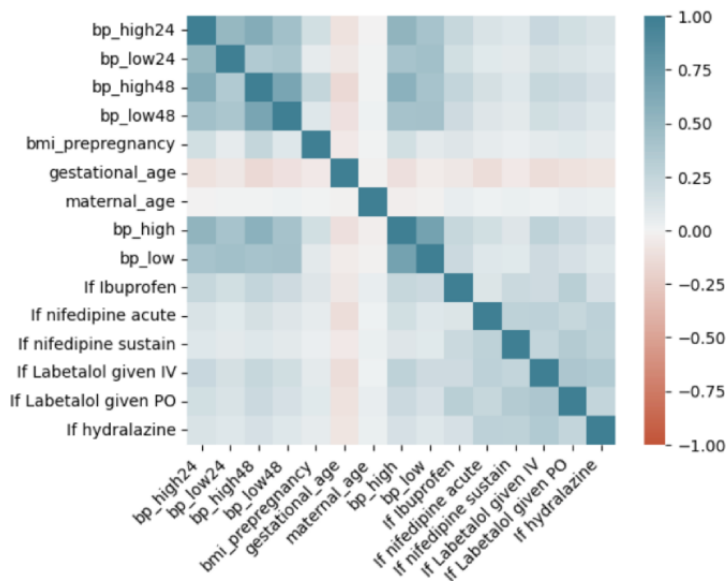
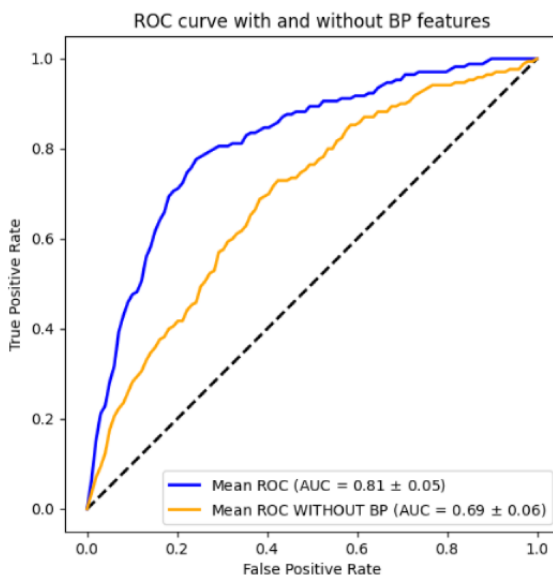


Figure 4. Receiver operating curve comparisons with and without blood pressure features including highest systolic blood pressure and associated diastolic blood pressure during 3 time periods—labor, between 0-24 hours postpartum, and between 24-48 hours postpartum. AUC: area under the curve.



Cost Analysis and Final Model Tuning

To tune and validate the final model deployed in a calculator, we also evaluated the model by measuring the estimated health care costs associated with the predictions. As previously mentioned, the value of a false negative was estimated to be US \$20,439, and the value of a false positive to be US \$36. The cost ratio was then created by dividing those 2 numbers and came out at 565 [17]. Lowering or raising this cost ratio places more weight on different factors; for example, the side effects associated with taking labetalol versus the time away from family or a job during a readmission. With this information, the estimated total cost for each model can be calculated by using the numbers of false negatives and false positives in the validation sets to give a sense for the medical impact of model implementation. However, there are a lot more factors that need

to be considered if we want the metric to be as generalizable as the balanced accuracy and the F_1 -score. For this reason, we did not consider estimated cost as the primary metric for model selection. In Table 3, we can see that of the models considered, the random forest model with class weight 1:200 had the highest balanced accuracy. Compared with the best logistic regression models and SVM, random forest with class weight 1:200 performs slightly better in terms of both balanced accuracy and F_1 -score. However, random forest model with class weight of 1:500 can provide better precision and F_1 -score. Combining the 2 metrics, we decided to implement the random forest model with class weight 1:500. Note that for all models the F_1 -score is relatively low, this is mainly due to the large imbalance in the data set. Since readmissions are fairly rare, to ensure that we avoid false negatives, we must reduce the precision of the model leading to the reduced score.

Table 3. Prediction model performance on joint data with cross-validation. For completeness, all models are included.

Candidate machine learning model	Specificity, %	Sensitivity, %	Precision or PPV ^a , %	NPV ^b , %	F_1 -score	Balanced accuracy, %	Cost (US \$)
Random forest model weights							
1	100	4.7	— ^c	99.5	—	52.35	659,016
200	77.2	78.2	1.77	99.85	0.0346	77.75 ^d	203,659.2
300	77.1	75.9	1.7	99.83	0.0332	76.5	220,284
500	79.1	74.1	1.82	99.83	0.0355 ^d	76.61	227,836.8
1000	79.2	71.7	1.77	99.81	0.0345	75.51	243,777.6
Decision tree model weights							
1	99.3	10	—	99.53	—	54.66	623,973.6
200	73.9	75.9	1.51	99.82	0.0296 ^d	74.87 ^d	227,908.8
300	75	72.4	1.5	99.81	0.0294	73.66	249,696
500	59.7	83.5	1.09	99.85	0.0215	71.62	208,080
1000	68.6	74.1	1.22	99.8	0.0240	71.35	252,446.4
Logistic regression (L2) model weights							
1	100	0	—	99.48	—	50	691,560
200	80.1	72.9	1.88	99.82	0.0366 ^d	76.53	233,596.8
300	71.6	83.5	1.51	99.88	0.0296	77.6 ^d	180,151.2
500	58.1	90.6	1.12	99.91	0.0221	74.36	162,964.8
1000	37.1	93.5	0.77	99.91	0.0152	65.33	191,757.6
Logistic regression (L1) model weights							
1	100	0	—	99.48	—	50	691,560
200	80.1	72.9	1.88	99.82	0.0366	76.52	233,625.6
300	71.6	83.5	1.51	99.88	0.0296	77.59 ^d	180,151.2
500	58.1	90.6	1.12	99.91	0.0221	74.36	162,964.8
1000	45.7	88.2	0.84	99.87	0.0166	66.99	208,180.8
SVM^e							
1	99.7	7	—	99.51	—	53.39	643,384.8
200	78.7	72.9	1.76	99.82	0.0343 ^d	75.84	236,800.8
300	71.1	84.1	1.5	99.88	0.0294	77.61 ^d	177,393.6
500	60.2	88.2	1.14	99.89	0.0225	74.21	174,456
1000	50	92.9	0.87	99.92	0.0172	68.96	177,429.6

^aPPV: positive predictive value.

^bNPV: negative predictive value.

^cNot available.

^dBest model with respect to the specific metric.

^eSVM: support vector machines.

Compared with the best logistic regression and SVM models, the random forest model with class weight 1:200 performs slightly better in terms of both balanced accuracy and F_1 -score. However, the random forest model with class weight 1:500 has a better precision and F_1 -score. Combining the 2 metrics, we decided to pick the random forest model with class weight 1:500

for final deployment. The overall hyperparameters picked for this model were a maximum tree depth of 6 and 100 total estimators.

With this model in mind, we performed a more in-depth cost analysis by varying the prediction threshold for the random forest model and examining how these impact the expected

medical costs and balanced accuracy. In Figures 5 and 6, we show the results of the analysis for expected costs and balanced accuracy, respectively. All values for this analysis were computed using leave-one-out cross-validation. As expected,

because the model was primarily chosen based on balanced accuracy, this measure is maximized at a threshold of 0.5. On the other hand, overall costs associated with model predictions are maximized at a threshold of 0.3.

Figure 5. Plot showing a relationship between predictive threshold for the model and expected medical costs associated with treatment based on model prediction. The y-axis is in thousands of US \$ and the x-axis represents the threshold of prediction. All values were computed using leave-one-out cross-validation and estimated costs from Niu et al [24].

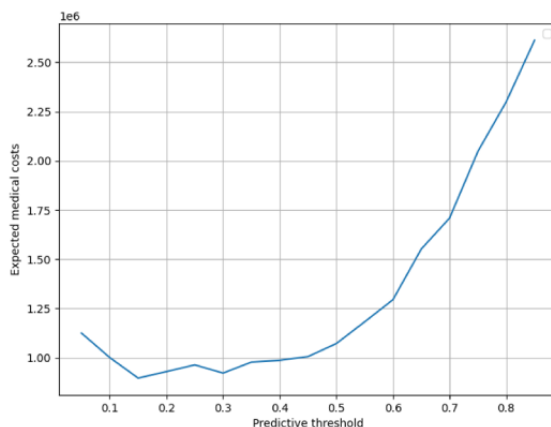
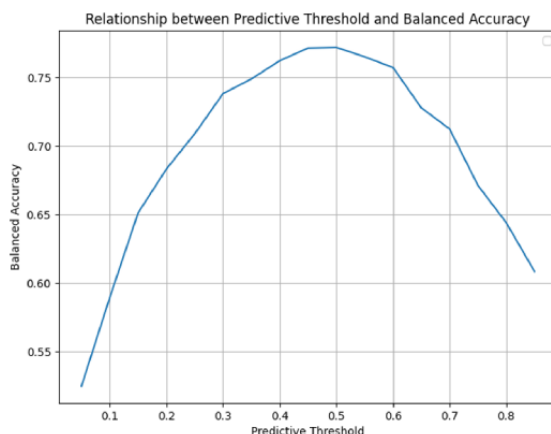


Figure 6. Plot showing a relationship between predictive threshold for the model and balanced accuracy associated with treatment based on model prediction. The y-axis is balanced accuracy and the x-axis represents the threshold of prediction. All values were computed using leave-one-out cross-validation and estimated costs from the other cost analysis.



Risk Calculator

We operationalized our model by incorporating it into a risk calculator that allows clinicians to compute how likely patients are to be readmitted for hypertension-related factors. Figure 7 shows a screenshot of our calculator; clinicians are able to enter 9 numerical features in text fields and click 6 binary features using a check box. The model for the calculator is deployed

using Python and Scikit-Learn [16] and is hosted on a public website. The full code of the calculator model is publicly available at the GitHub repository [19]. Based on the results from our previous analysis, in practice clinicians using this tool may want to consider any likelihood above 30% as important to consider when making treatment decisions to minimize readmission risk and related medical costs.

Figure 7. Screenshot of calculator website.

Discussion

Principal Results

Our research indicates the importance of intrapartum and postpartum blood pressure measurements in predicting readmission. This is clinically important, as it suggests blood pressure metrics before birth may be more important in guiding postpartum hypertension treatment than previously acknowledged. Current management of hypertension in pregnancy is based on expert opinion and has recommended initiation of antihypertensive medication for postpartum systolic blood pressure greater than 150 mm Hg or diastolic blood pressure greater than 100 mm Hg [4]. A recent study published in *AJOG MFM (American Journal of Obstetrics & Gynecology Maternal-Fetal Medicine)* suggests that lowering this threshold to 140 mm Hg systolic or 90 mm Hg diastolic can increase sensitivity in predicting postpartum readmission [10]. Regardless, given that systemic vascular resistance remains at the pregnancy-associated lower value for about 2 days and then subsequently increases to normal prepregnancy values by postpartum day 3 to day 4, many women may be discharged before the postpartum equilibration of blood pressure on postpartum day 3 to day 4 and thus may be undertreated [20-23]. Using peak blood pressure values obtained during labor to aid in decision-making may improve triaging and treatment of hypertension after delivery, thus decreasing the risk of postpartum readmission. Our research additionally indicates that blood pressure metrics themselves are more important in predicting readmission than more typically used patient demographics such as gestational age at delivery, maternal age, BMI, laboratory data, or the administration of oral or intravenous antihypertensive medication before discharge [11]. Perhaps the awareness of more severe diseases allows for more aggressive treatment, management, and follow-up after the initial hospital stay, thus decreasing readmission rates in this higher-risk group.

Comparison With Previous Work

Research evaluating postpartum readmission has used descriptive statistics to describe demographic variables

implicated in readmission. A nested case-control study published in the *Journal of Perinatology* in 2016 demonstrated no increased risk of readmission by mode of delivery, severity of preeclampsia, fluid balance, use of magnesium sulfate, or lab abnormalities but did find a decreased risk of readmission for women discharged home on antihypertensive medication when controlled for age, race, and presence of chronic hypertension [24]. However, this study only included women with hypertension during their initial labor and delivery admission. Given that 30% of women who experience hypertension-related postpartum readmission do not have antecedent diagnoses of hypertension, we furthered the previously published work by Hirshberg et al [24] and included women without known prepregnancy or pregnancy-induced hypertension in evaluating for postpartum readmission in this study. Recently, an ML model was published that evaluated factors predictive of hypertension-related postpartum readmission [11]. Hoffman et al [11] evaluated 31 features in their model, similarly finding that systolic blood pressure (specifically the moving average, or trend of the systolic blood pressure) was the most important predictor of readmission. Our model identified biometric, demographic, and obstetric variables easily identified in any patient's medical record. In addition, we investigated the use of specific antihypertensive medication, rather than using a drug score that does not indicate which specific agents were used. We used a cost-sensitive random forest method, allowing us to weigh the importance of particular observations and thus penalize false negatives significantly higher than false positives.

Using our data and findings from this analysis, we created a clinical risk calculator [25]

that predicts the likelihood of readmission based upon the key clinical variables found to be most predictive of hypertension-related postpartum readmission. Similar risk-based calculators have previously been created and validated, including the vaginal birth after cesarean calculator, commonly used during the antepartum period to guide counseling and management of women with a previous cesarean section, and more recently a calculator to estimate the risk of cesarean section after an induction of labor with an unfavorable cervix [26,27].

Our calculator applies our predictive model to any given patient to predict the likelihood of readmission. While we do not define when a patient should or should not be treated based on the likelihood of readmission, we hope that better quantifying the likelihood of readmission will allow for an improved discussion between health care providers and their patients. However, based on our cost analysis, it seems that if the calculator reports a likelihood of 30% or higher, clinicians may want to seriously consider treatment options to reduce costs related to readmission. Management options for women at higher risk of readmission include earlier initiation of antihypertensive medication or closer outpatient blood pressure surveillance with daily remote patient monitoring or self-monitoring. These interventions would hopefully lead to decreased health care costs by transitioning to outpatient rather than inpatient care models. We recognize that a balanced accuracy of 76.61% allows for error in our model, thus health care providers must include this in their counseling to optimize shared decision-making. Also, the 1.82% precision that allows a number of false alarms should be noted. However, since the cost of misidentifying a readmission is quite high, the rate of false alarms might be necessary to provide adequate care in the absence of other treatment options such as home monitoring.

Strengths and Limitations

Strengths of our research include the development of a predictive model, different from the previously used descriptive models. In addition, we had a large data set comprised from several sources allowing for better validation and model development. One limitation of our research is that the proposed predictive model is a random forest method, which is difficult to interpret. Unlike logistic regression or a single decision tree, it is difficult to extract exact thresholds for particular clinical measurements to determine how they will impact the output of the random forest. The importance plot of the random forest can be used to find which features are most important for the model to make a prediction, but it cannot be used to determine particular

prediction thresholds. This is of particular importance as these thresholds will be key in establishing new treatment protocols. In order to retrieve such thresholds, additional research needs to be done in extracting an explainer model from our random forest. Additional limitations include the use of *ICD-9-CM (International Classification of Diseases, Ninth Revision, Clinical Modification)* and *ICD-10 (International Statistical Classification of Diseases, Tenth Revision)* codes for diagnosis of preexisting hypertension and HDP, which may have led to errors in coding and underreporting. Our rate of hypertension was 9%, which aligns with national data, but is lower than that previously reported in Wisconsin (estimated at 22%). Finally, our readmission rate was low at <1%. Given the large catchment area of the institution used in our research, it is possible that women who delivered at our hospital presented to their local emergency department or provider with postpartum hypertension and were thus not included in our data as a readmit. This study is additionally limited by generalizability. Our patients came from a single, relatively homogenous, midwestern academic institution. In order to apply these findings more broadly, our predictive model should be applied to a more diverse population.

Conclusions

Our research shows that blood pressure metrics during labor and post partum, in addition to obstetric and demographic variables, are critical in creating a predictive model for postpartum readmission. Predictive models like ours can improve postpartum management, allowing practitioners to characterize women as low-risk and high-risk for readmission and better individualize treatment. If we can better predict readmission, we can better prevent readmission. By creating a clinical calculator to help guide postpartum hypertension treatment, our goal is to decrease adverse maternal outcomes and prevent costly postpartum readmission. Future research will involve validating this model and finding specific threshold values at which treatment is to be initiated.

Acknowledgments

The authors would also like to thank Mr Dakota Dalton for his help with the deployment of the web-based calculator. The research obtained financial support from the University of Wisconsin Department of Obstetrics & Gynecology Research and Development. KKH was supported by the ObGyn Start-Up-UWF-SMPH Research and Development-AAB8581 award. No generative artificial intelligence was used in the creation of the manuscript.

Data Availability

The data sets generated during or analyzed during this study are not publicly available as we do not standardly post or publish data sets used for retrospective clinical research projects, and we also do not have Institutional Review Board's approval to post the data set publicly. However, the data set is available from the corresponding author on reasonable request.

Authors' Contributions

RGL wrote the initial draft of the manuscript and performed data curation through chart review. JT performed formal analysis of the data, validated the results, aided in the revisions, and worked on the software deployment of the online calculator. YM supervised JT, created the core methodology for the analysis in the paper, and performed project administration. He also contributed to writing the edited draft. OA assisted with methodology creation and writing of the original draft. KKH conceptualized this project idea, assisted with data collection, supervised RGL, and oversaw the research project.

Conflicts of Interest

OA served as a paid consultant for Bristol Myers Squibb, Johnson & Johnson, and Exact Sciences outside the submitted work; and is the owner of and principal scientist for Innovo Analytics LLC outside the submitted work.

Multimedia Appendix 1

Additional figures and tables, and supplementary analysis.

[[DOCX File , 1894 KB - ai_v3i1e48588_app1.docx](#)]

References

1. Roberts CL, Algert CS, Morris JM, Ford JB, Henderson-Smart DJ. Hypertensive disorders in pregnancy: a population-based study. *Med J Aust* 2005;182(7):332-335. [doi: [10.5694/j.1326-5377.2005.tb06730.x](https://doi.org/10.5694/j.1326-5377.2005.tb06730.x)] [Medline: [15804223](#)]
2. Clapp MA, Little SE, Zheng J, Kaimal AJ, Robinson JN. Hospital-level variation in postpartum readmissions. *JAMA* 2017;317(20):2128-2129 [FREE Full text] [doi: [10.1001/jama.2017.2830](https://doi.org/10.1001/jama.2017.2830)] [Medline: [28535223](#)]
3. Clapp MA, Little SE, Zheng J, Robinson JN. A multi-state analysis of postpartum readmissions in the United States. *Am J Obstet Gynecol* 2016;215(1):113.e1-113.e10. [doi: [10.1016/j.ajog.2016.01.174](https://doi.org/10.1016/j.ajog.2016.01.174)] [Medline: [27829570](#)]
4. NA. Hypertension in pregnancy. Report of the American college of obstetricians and gynecologists' task force on hypertension in pregnancy. *Obstet Gynecol* 2013;122(5):1122-1131. [doi: [10.1097/01.AOG.0000437382.03963.88](https://doi.org/10.1097/01.AOG.0000437382.03963.88)] [Medline: [24150027](#)]
5. Parikh NI, Laria B, Nah G, Singhal M, Vittinghoff E, Vieten C, et al. Cardiovascular disease-related pregnancy complications are associated with increased maternal levels and trajectories of cardiovascular disease biomarkers during and after pregnancy. *J Womens Health (Larchmt)* 2020;29(10):1283-1291 [FREE Full text] [doi: [10.1089/jwh.2018.7560](https://doi.org/10.1089/jwh.2018.7560)] [Medline: [31934809](#)]
6. Parikh NI, Norberg M, Ingelsson E, Cnattingius S, Vasani RS, Domellöf M, et al. Association of pregnancy complications and characteristics with future risk of elevated blood pressure: the västerbotten intervention program. *Hypertension* 2017;69(3):475-483 [FREE Full text] [doi: [10.1161/HYPERTENSIONAHA.116.08121](https://doi.org/10.1161/HYPERTENSIONAHA.116.08121)] [Medline: [28137991](#)]
7. Kaaja RJ, Greer IA. Manifestations of chronic disease during pregnancy. *JAMA* 2005;294(21):2751-2757. [doi: [10.1001/jama.294.21.2751](https://doi.org/10.1001/jama.294.21.2751)] [Medline: [16333011](#)]
8. Fraser A, Nelson SM, Macdonald-Wallis C, Cherry L, Butler E, Sattar N, et al. Associations of pregnancy complications with calculated cardiovascular disease risk and cardiovascular risk factors in middle age: the Avon Longitudinal Study of Parents and Children. *Circulation* 2012;125(11):1367-1380 [FREE Full text] [doi: [10.1161/CIRCULATIONAHA.111.044784](https://doi.org/10.1161/CIRCULATIONAHA.111.044784)] [Medline: [22344039](#)]
9. Mogos MF, Salemi JL, Spooner KK, McFarlin BL, Salihi HH. Hypertensive disorders of pregnancy and postpartum readmission in the United States: national surveillance of the revolving door. *J Hypertens* 2018;36(3):608-618. [doi: [10.1097/HJH.0000000000001594](https://doi.org/10.1097/HJH.0000000000001594)] [Medline: [29045342](#)]
10. Mukhtarova N, Alagoz O, Chen YH, Hoppe K. Evaluation of different blood pressure assessment strategies and cutoff values to predict postpartum hypertension-related readmissions: a retrospective cohort study. *Am J Obstet Gynecol MFM* 2021;3(1):100252. [doi: [10.1016/j.ajogmf.2020.100252](https://doi.org/10.1016/j.ajogmf.2020.100252)] [Medline: [33451628](#)]
11. Hoffman MK, Ma N, Roberts A. A machine learning algorithm for predicting maternal readmission for hypertensive disorders of pregnancy. *Am J Obstet Gynecol MFM* 2021;3(1):100250. [doi: [10.1016/j.ajogmf.2020.100250](https://doi.org/10.1016/j.ajogmf.2020.100250)] [Medline: [33451620](#)]
12. PeriData. URL: <https://www.peridata.net/peridata/security/access.htm> [accessed 2024-08-14]
13. McKinney W. Data structures for statistical computing in python. 2010 Presented at: Proceedings of the 9th Python in Science Conference; 2010 Jun 28; Austin, Texas p. 51-56. [doi: [10.25080/majora-92bf1922-00a](https://doi.org/10.25080/majora-92bf1922-00a)]
14. Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. 2nd ed. New York: Springer-Verlag; 2009:9780387848587.
15. Tharwat A. Classification assessment methods. *ACI* 2020;17(1):168-192. [doi: [10.1016/j.aci.2018.08.003](https://doi.org/10.1016/j.aci.2018.08.003)]
16. Pedregosa F, Varoquaux G, Gramfort A, Michel V. Scikit-learn: machine learning in python. *Journal of Machine Learning Research* 2011;12:2825-2830.
17. Niu B, Mukhtarova N, Alagoz O, Hoppe K. Cost-effectiveness of telehealth with remote patient monitoring for postpartum hypertension. *J Matern Fetal Neonatal Med* 2022;35(25):7555-7561. [doi: [10.1080/14767058.2021.1956456](https://doi.org/10.1080/14767058.2021.1956456)] [Medline: [34470135](#)]
18. Hypertension prediction. URL: <https://github.com/jtao34/hypertension-prediction> [accessed 2024-08-14]
19. Hypertension calculator. URL: <https://github.com/uwobgyn/HypertensionCalculator> [accessed 2024-08-14]
20. Hibbard JU, Schroff SG, Cunningham FG. Cardiovascular alterations in normal and preeclamptic pregnancy. In: Taylor RN, Roberts JM, Cunningham FG, editors. *Chesley's Hypertensive Disorders in Pregnancy (Fourth Edition)*. US: Academic Press; 2014:291-313.
21. Christianson RE. Studies on blood pressure during pregnancy. I. influence of parity and age. *Am J Obstet Gynecol* 1976;125(4):509-513. [doi: [10.1016/0002-9378\(76\)90367-7](https://doi.org/10.1016/0002-9378(76)90367-7)] [Medline: [984086](#)]
22. Cunningham FG, Leveno KJ, Bloom SL, Spong CY, Dashe JS, Hoffman BL, et al. In: *Obstetrics W*, editor. *The Puerperium*. US: McGraw-Hill Education; 2014:668-681.

23. Walters BN, Thompson ME, Lee A, de Swiet M. Blood pressure in the puerperium. *Clin Sci (Lond)* 1986;71(5):589-594. [doi: [10.1042/cs0710589](https://doi.org/10.1042/cs0710589)] [Medline: [3769407](https://pubmed.ncbi.nlm.nih.gov/3769407/)]
24. Hirshberg A, Levine LD, Srinivas SK. Clinical factors associated with readmission for postpartum hypertension in women with pregnancy-related hypertension: a nested case control study. *J Perinatol* 2016;36(5):405-409 [FREE Full text] [doi: [10.1038/jp.2015.209](https://doi.org/10.1038/jp.2015.209)] [Medline: [26765549](https://pubmed.ncbi.nlm.nih.gov/26765549/)]
25. Staying Healthy After Childbirth.: Department of Obstetrics and Gynecology; 2023. URL: <https://www.obgyn.wisc.edu/stac/ObGyn.HypertensionCalculator> [accessed 2024-07-01]
26. Grobman WA, Lai Y, Landon MB, Spong CY, Leveno KJ, Rouse DJ, National Institute of Child Health Human Development (NICHD) Maternal-Fetal Medicine Units Network (MFMU). Development of a nomogram for prediction of vaginal birth after cesarean delivery. *Obstet Gynecol* 2007;109(4):806-812. [doi: [10.1097/01.AOG.0000259312.36053.02](https://doi.org/10.1097/01.AOG.0000259312.36053.02)] [Medline: [17400840](https://pubmed.ncbi.nlm.nih.gov/17400840/)]
27. Levine LD, Downes KL, Parry S, Elovitz MA, Sammel MD, Srinivas SK. A validated calculator to estimate risk of cesarean after an induction of labor with an unfavorable cervix. *Am J Obstet Gynecol* 2018;218(2):254.e1-254.e7 [FREE Full text] [doi: [10.1016/j.ajog.2017.11.603](https://doi.org/10.1016/j.ajog.2017.11.603)] [Medline: [29224730](https://pubmed.ncbi.nlm.nih.gov/29224730/)]

Abbreviations

AJOG MFM: American Journal of Obstetrics & Gynecology Maternal-Fetal Medicine

EMR: electronic medical record

HDP: hypertensive disorders of pregnancy

ICD-9-CM: International Classification of Diseases, Ninth Revision, Clinical Modification

ICD-10: International Statistical Classification of Diseases, Tenth Revision

ML: machine learning

ROC: receiver operating characteristic

SVM: support vector machines

Edited by K El Emam, B Malin; submitted 28.04.23; peer-reviewed by J Luo, S Sarejloo, T Wang, T Kotani; comments to author 02.01.24; revised version received 07.02.24; accepted 23.06.24; published 13.09.24.

Please cite as:

Tao J, Larson RG, Mintz Y, Alagoz O, Hoppe KK

Predictive Modeling of Hypertension-Related Postpartum Readmission: Retrospective Cohort Analysis

JMIR AI 2024;3:e48588

URL: <https://ai.jmir.org/2024/1/e48588>

doi: [10.2196/48588](https://doi.org/10.2196/48588)

PMID:

©Jinxin Tao, Ramsey G Larson, Yonatan Mintz, Oguzhan Alagoz, Kara K Hoppe. Originally published in JMIR AI (<https://ai.jmir.org>), 13.09.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR AI, is properly cited. The complete bibliographic information, a link to the original publication on <https://www.ai.jmir.org/>, as well as this copyright and license information must be included.

Original Paper

Identifying Patterns of Smoking Cessation App Feature Use That Predict Successful Quitting: Secondary Analysis of Experimental Data Leveraging Machine Learning

Leeann Nicole Siegel¹, MPH, PhD; Kara P Wiseman², MPH, PhD; Alex Budenz¹, MA, DrPH; Yvonne Prutzman¹, MPH, PhD

¹National Cancer Institute, National Institutes of Health, Rockville, MD, United States

²University of Virginia School of Medicine, Charlottesville, VA, United States

Corresponding Author:

Kara P Wiseman, MPH, PhD

University of Virginia School of Medicine

PO Box 800717

Charlottesville, VA, 22908

United States

Phone: 1 4342438126

Email: kara.wiseman@virginia.edu

Abstract

Background: Leveraging free smartphone apps can help expand the availability and use of evidence-based smoking cessation interventions. However, there is a need for additional research investigating how the use of different features within such apps impacts their effectiveness.

Objective: We used observational data collected from an experiment of a publicly available smoking cessation app to develop supervised machine learning (SML) algorithms intended to distinguish the app features that promote successful smoking cessation. We then assessed the extent to which patterns of app feature use accounted for variance in cessation that could not be explained by other known predictors of cessation (eg, tobacco use behaviors).

Methods: Data came from an experiment (ClinicalTrials.gov NCT04623736) testing the impacts of incentivizing ecological momentary assessments within the National Cancer Institute's quitSTART app. Participants' (N=133) app activity, including every action they took within the app and its corresponding time stamp, was recorded. Demographic and baseline tobacco use characteristics were measured at the start of the experiment, and short-term smoking cessation (7-day point prevalence abstinence) was measured at 4 weeks after baseline. Logistic regression SML modeling was used to estimate participants' probability of cessation from 28 variables reflecting participants' use of different app features, assigned experimental conditions, and phone type (iPhone [Apple Inc] or Android [Google]). The SML model was first fit in a training set (n=100) and then its accuracy was assessed in a held-aside test set (n=33). Within the test set, a likelihood ratio test (n=30) assessed whether adding individuals' SML-predicted probabilities of cessation to a logistic regression model that included demographic and tobacco use (eg, polyuse) variables explained additional variance in 4-week cessation.

Results: The SML model's sensitivity (0.67) and specificity (0.67) in the held-aside test set indicated that individuals' patterns of using different app features predicted cessation with reasonable accuracy. The likelihood ratio test showed that the logistic regression, which included the SML model-predicted probabilities, was statistically equivalent to the model that only included the demographic and tobacco use variables ($P=.16$).

Conclusions: Harnessing user data through SML could help determine the features of smoking cessation apps that are most useful. This methodological approach could be applied in future research focusing on smoking cessation app features to inform the development and improvement of smoking cessation apps.

Trial Registration: ClinicalTrials.gov NCT04623736; <https://clinicaltrials.gov/study/NCT04623736>

(JMIR AI 2024;3:e51756) doi:[10.2196/51756](https://doi.org/10.2196/51756)

KEYWORDS

smartphone apps; machine learning; artificial intelligence; smoking cessation; mHealth; mobile health; app; apps; applications; application feature; features; smoking; smoke; smoker; smokers; cessation; quit; quitting; algorithm; algorithms; mobile phone

Introduction

Cigarette smoking remains a leading cause of preventable death in the United States [1]. Evidence-based smoking cessation interventions, though proven to be valuable in helping people quit, are underused [2]. Smartphone apps have the potential to expand the reach and increase the use of evidence-based smoking cessation interventions [1,3]. Smartphone ownership is high among every demographic group in the United States [4], and an array of smoking cessation apps, including many free options [5], are available in smartphone app stores. Evidence suggests that smoking cessation apps are widely used, with millions of downloads per year [6,7].

Research supporting the use of apps for smoking cessation is still emerging, and many publicly available apps have not been rigorously tested [8]. However, results from randomized controlled trials (RCTs) suggest that apps can be effective in helping people quit smoking [9-11]. Studies have also demonstrated that both higher user engagement in smoking cessation apps [11,12] and longer duration and greater frequency of app use [6] are related to smoking cessation.

The many capabilities, features, and functionalities that can be incorporated into smoking cessation apps have the potential to increase their effectiveness. Apps can include interactive and multimedia content, and offer tailored features to meet the needs and preferences of different types of users [13]. Several reviews have cataloged the most common types of features in smoking cessation apps and evaluated whether those features align with behavioral theories or smoking cessation clinical guidelines [5,14-17]. Some studies have also investigated whether and how users respond to and use particular app features. Through their content analysis of smoking cessation app reviews and ratings, Bendotti et al [18] found that users liked app features that allowed them to set goals, track their progress, understand and manage their cigarette cravings, and interact with others within the app. Hoepfner et al [13] found that apps using tailored communications with users were more likely to have received more than 10,000 downloads compared to apps that did not use tailored communications. In a recent study focused on the National Cancer Institute's quitSTART app, the app used in this study, Budenz et al [19] found that a substantial proportion of users accessed app-integrated, mood-related support.

Few studies have examined the impacts of using particular app features on smoking cessation outcomes. Rajani et al [20] found that increased frequency of use of their apps' gamification features (eg, earning badges and unlocking levels) was associated with increases in perceived self-efficacy and motivation to quit smoking. Heffner et al [21] looked at features within a smoking cessation app that was both popular (ie, among the 10 most-used features in the app) and significantly associated with successful quitting and identified 2 app features that met both criteria—viewing one's quit plan and tracking one's

practice of letting smoking urges pass. In their study focused on a smoking cessation app that emphasized positive psychology content, Hoepfner et al [22] found that greater engagement with the app's happiness-related features was predictive of cessation.

More research is needed to understand which smoking cessation app features are most valuable in helping users quit smoking. Fortunately, the apps are designed to efficiently collect user data that can be used to answer this question. App developers can record users' activity within apps, capturing information such as how many times and when an individual took an action within the app and how quickly they responded to an app notification. However, raw app user data can be large and unwieldy, particularly for apps that offer many features and garner frequent engagement from users. Machine learning methods expand our ability to analyze and glean insights from app user data. The use of machine learning methods to analyze user data from smoking cessation apps has the potential to optimize the effectiveness of such apps [23,24].

In this study, we leverage supervised machine learning (SML) methods to conduct a secondary analysis of app user data collected from participants as part of an RCT involving the quitSTART smoking cessation app—the quitSTART-EMA Incentivization Trial. Our primary goal in conducting this study is to outline an analytic approach that could be used in future studies investigating whether and how patterns of use of different smoking cessation app features affect cessation. We also seek to fulfill the following exploratory research aims: (1) examine the extent to which patterns of use of different features of the quitSTART app can be used to predict participants' short-term smoking cessation and (2) test whether participants' patterns of app feature use predict variance in short-term cessation that is not predicted by other variables related to smoking cessation.

Methods

The quitSTART App

The quitSTART app is a free, publicly available app created by the National Cancer Institute's Smokefree.gov initiative, a federal program that offers no-cost, evidence-based tobacco cessation support to the public through a suite of websites, text messaging programs, and mobile apps [25]. The quitSTART app is available for both iPhones and Androids and is popular, with 10,000-20,000 new downloads each year [25].

The app offers a range of features designed to assist individuals in quitting smoking. App users can explore content pages, referred to as "cards," which contain information, tips, and inspiration for quitting smoking. They can also seek real-time support for managing their cravings, mood, and handling slips; play games to distract themselves during cravings; track their progress; and earn badges as they continue to use the app. Users can customize their app experience by building a "quit kit" containing cards they find useful and can create custom

notifications. Since 2017, the quitSTART app has also included ecological momentary assessment (EMA) capability. By default, users are sent 3 EMA prompts each day at random times to report their craving level, mood, and number of cigarettes smoked. Users can opt out of receiving EMAs by disabling notifications from the app.

Experimental Design

Data for this analysis were drawn from an experimental trial conducted between October 2020 and May 2021. The quitSTART EMA Incentivization Trial was conducted to test the effects of incentivizing EMA completion within the quitSTART app on short-term smoking cessation. Participants were English-speaking adults who lived in the United States, smoked cigarettes, and had a self-reported desire to quit smoking.

As the goal of the clinical trial was to test the effects of incentivizing completion of EMAs on smoking cessation, eligible participants were randomized 1:1 into 2 study arms, an incentivized EMA arm and a nonincentivized EMA arm. Participants randomized to the nonincentivized EMA arm were compensated for completing the surveys administered to all participants at baseline, 2 weeks into the study, and at the end of the 4-week study. Participants in the nonincentivized arm received EMA notifications, which are sent to all users by default. However, their compensation was not affected by their EMA completion. In contrast, participants randomized to the incentivized EMA arm were informed that part of their compensation would be contingent on completing surveys and the other part would be contingent on their EMA participation. They had to complete at least half of the programmed EMAs to receive any EMA compensation, and increasing EMA participation resulted in higher compensation. The total amount of compensation that could be earned was identical across the 2 study arms.

After completing the baseline survey, participants were instructed to download the quitSTART app and use it for the 4-week study period. A total of 152 participants completed the enrollment process and participated in the study, of whom 133 (88.2%) completed the 4-week follow-up survey. These 133 participants were included in this study. Figure S1 in [Multimedia Appendix 1](#) summarizes the recruitment, randomization, and data collection processes for this study.

Ethical Considerations

The University of Virginia institutional review board approved the study design and protocol (UVA SBS IRB protocol 3643; ClinicalTrials.gov NCT04623736).

Study Measures

Baseline Participant Characteristics

Data collected in the baseline survey included participants' gender identity, sexual orientation, education level, and scores on the Patient Health Questionnaire-9 (PHQ-9) [26], which is used to measure the presence and severity of depressive symptoms. The baseline survey also assessed participants' use of tobacco products, nicotine dependence scores [27], and whether they had made an attempt to quit within the past year.

When participants downloaded and first used the app, their phone type (ie, whether they had an Android or iPhone) was recorded.

Smoking Cessation Outcome Measure

The outcome of interest for this study, short-term cigarette smoking cessation, was measured at the end of the quitSTART EMA Incentivization Trial and was operationalized as 7-day point-prevalence abstinence at 4 weeks postenrollment. Participants were asked, "Have you smoked a cigarette (even a puff) in the past seven days?" Participants who responded "no" to this question were considered to have quit smoking.

App Feature Use Variables

As participants used the quitSTART app, each action they took and its corresponding time stamp were recorded. These data were used to create 3 sets of variables reflecting the participants' use of app features. The first set of variables, "binary app feature use variables," consisted of yes or no variables that reflected whether a participant took the action in question; these variables were used for actions that most participants took only 1 time (eg, completing the initial profile set-up process).

For actions that were intended to be taken as many times as a participant wanted (eg, playing a game), 2 additional sets of variables were created. One set of variables used in our main analyses, which we labeled "proportion app feature use variables," reflected the number of times a participant took a particular action within the app divided by their total number of app use sessions. An app use session was defined as a period during which a participant performed 1 or more actions in the app with no more than 2 minutes between actions. We took this approach to ensure that we captured variation in how participants spent their time within the app rather than just variation in the total time they spent in the app. The other set of variables, "count app feature use variables," reflected the total number of times participants took an action and were used in a sensitivity analysis, as described below.

Data Analysis

Overview

All analyses were conducted in R (version 4.1; R Core Team). We first examined descriptive statistics for the baseline participant characteristics. We also examined participants' responses to our short-term smoking cessation item.

Our machine learning approach was based on the recommendations made by Dinga et al [28] for controlling for the effects of confounding variables on machine learning predictions. Dinga et al [28] argued that regressing out confounding variables from each predictor variable separately prior to conducting machine learning modeling is insufficient. They instead proposed controlling for confounding variables post hoc at the level of machine learning predictions. We adopted this approach for 2 reasons. First, it allowed us to control for confounding more efficiently. It also enabled us to fulfill our second study aim by testing whether predictions from our machine learning model, which included input variables capturing participants' use of different app features, explained variance in cessation that was not explained by participant-level

variables that could potentially affect cessation such as demographic characteristics and tobacco use.

Aim 1 Analysis

To identify which patterns of use of app feature use predict short-term smoking cessation, we built SML models predicting 7-day smoking abstinence from a set of predictor variables that included our binary app feature use variables, our proportion app feature use variables, participants' total number of app use sessions, phone type (iPhone or Android), and study arm. Phone type was included as a variable in the SML models because the iPhone and Android versions of the quitSTART app were built separately and user data from each app were recorded in a slightly different manner. Although the 2 apps appeared identical to users and we harmonized the user data collected from each, we chose to include phone type as a variable in the SML models in case there was a relationship between phone type and app use or between phone type and cessation. We first randomly divided our data into a training set ($n=100$, 75% of the data) and a held-aside test set ($n=33$, 25% of the data).

Working with the training set, we used recursive feature elimination with 10-fold cross-validation to determine the optimal number of features for our classifier and then fit our logistic regression classifier using this number of features to the training set. We selected a logistic regression classifier because our outcome variable was binary, and we wanted a classifier that would yield predicted probabilities (rather than binary predictions) for every participant. We evaluated the SML model's performance in the training set by looking at its sensitivity, specificity, and accuracy. We also examined variable importance (defined as the scaled absolute value of the coefficient of each variable in a logistic regression model for binary classification) for each feature and identified the features in the model assigned the highest importance for predicting cessation. We produced partial dependence plots for each of the top 10 most important features in order to better understand each feature's relationship with short-term smoking cessation [29].

We then applied the model to the held-aside test set and looked at its sensitivity, specificity, and accuracy. We then used it to produce predicted probabilities of short-term cessation for each participant included in the test set.

Aim 2 Analysis

As a first step toward testing whether participants' patterns of app feature use predicted unique variance in cessation, we fit 2 logistic regression models using the test set data. These models were fit with all participants in the test set who were not missing data on any demographic or participant characteristic variables ($n=30$; a total of 3 participants were excluded from the aim 2 analyses because of missing data on the gender variable). Due to the small sample size available, no data splitting or cross-validation was performed. The first model included participant demographic variables, as well as other variables

that prior research suggests may be related to cessation. These variables were measured in the baseline survey and included age, race or ethnicity, gender identity, education, PHQ-9 scores, sexual orientation, nicotine dependence, quit attempts in the past year, and polytobacco use. The second model included all these variables, as well as an additional predictor variable—the predicted probabilities of short-term cessation from the SML model. After fitting each model, we assessed its fit through a likelihood ratio test comparing it to a null model. We then ran a likelihood ratio test comparing the 2 logistic regression models to one another to assess whether the model that included the SML model-predicted probabilities of cessation had a significantly better fit to the data.

Sensitivity Analysis

As a sensitivity analysis, we repeated our aim 1 and aim 2 analyses with 1 major change. We used the count app feature use variables in place of the proportion app feature use variables in our SML model. Participants' total number of app use sessions was not included as a predictor in these models due to its collinearity with the count app feature use variables.

Results

Descriptive statistics for participant characteristics measured in the baseline survey, as well as participants' study arm and phone type, are summarized in [Table 1](#). Descriptive statistics are shown for all participants, as well as for participants who were included in the training set ($n=100$) and in the test set ($n=33$) when building our SML models. Among all 133 participants in the study, 62 (46.6%) were randomized to the incentivized EMA arm. About half ($n=74$, 55.6%) of participants had iPhones, while 59 (44.4%) had Androids. Participants' average age was 45.6 (SD 12.6) years. Participants reported being mostly non-Hispanic White ($n=103$, 77.4%), female ($n=99$, 74.4%), and straight ($n=106$, 79.7%). The average PHQ-9 score was 7.8 (SD 6.1), which indicates mild depression [26].

Participants' mean score on the Fagerstrom test was 4.8 (SD 2.4), which equates to medium nicotine dependence [30]. Most participants ($n=105$, 78.9%) had made a prior attempt to quit smoking within the past year. Approximately a third of participants ($n=46$, 34.6%) reported polytobacco use. Roughly a quarter ($n=37$, 27.8%) of participants reported 7-day point-prevalence abstinence at 4 weeks.

The full list of variables that were considered for inclusion in the SML model and their descriptions are included in [Table 2](#). Results from recursive feature elimination showed that 28 features out of 29 candidate features should be included in the SML model (every feature except `ncravingspressed_prop`). We ran our SML model including these 28 features in the training set and assessed its performance. The model's accuracy in the training set was 0.91, its sensitivity was 0.96, and its specificity was 0.79.

Table 1. Baseline participant characteristics for all participants, training set, and test set.

Characteristics	All participants (N=133)	Training set (n=100)	Test set (n=33)
Study arm, n (%)			
Incentivized EMA ^a arm	62 (46.6)	45 (45)	17 (51.5)
Nonincentivized EMA arm	71 (53.4)	55 (55)	16 (48.5)
Phone type, n (%)			
iPhone	74 (55.6)	53 (53)	21 (63.6)
Android	59 (44.4)	47 (47)	12 (36.4)
Age (years), mean (SD)	45.6 (12.6)	47.0 (12.4)	41.5 (12.3)
Race or ethnicity, n (%)			
Non-Hispanic White	103 (77.4)	77 (77)	26 (78.8)
Hispanic White	30 (22.6)	23 (23)	7 (21.2)
Sex, n (%)			
Male	31 (23.3)	25 (25)	6 (18.2)
Female	99 (74.4)	75 (75)	24 (77.4)
Missing	3 (2.3)	0 (0)	3 (9.1)
Education level, n (%)			
Less than high school	6 (4.5)	5 (5)	1 (3)
High school graduate or equivalent	11 (8.3)	8 (8)	3 (9.1)
Some college	50 (37.6)	37 (37)	13 (39.4)
College graduate or more	66 (49.6)	50 (50)	16 (48.5)
Sexual minority status, (%)			
Straight	106 (79.7)	84 (84)	22 (66.7)
Not straight	27 (20.3)	16 (16)	11 (33.3)
PHQ-9 ^b score, mean (SD)	7.8 (6.1)	8.07 (6.32)	7.15 (5.59)
Fagerstrom test, mean (SD)	4.8 (2.4)	4.56 (2.31)	5.52 (2.55)
Quit attempt in past 12 months, n (%)			
Yes	105 (78.9)	78 (78)	27 (81.8)
No	28 (21.1)	22 (22)	6 (18.2)
Poly-use of tobacco products, n (%)			
Yes	46 (34.6)	34 (34)	12 (36.4)
No	87 (65.4)	66 (66)	21 (63.6)

^aEMA: ecological momentary assessment.

^bPHQ-9: Patient Health Questionnaire-9.

Table 2. Variables considered for inclusion in SML^a model (n=29), definitions, and mean values among participants (N=133).

Variable name	Definition	Values
Proportion app feature use variables (n=24), mean (SD)		
naddlocation_prop	How many times a participant added a location to receive a location-based notification app use divided by their app use sessions.	0.02 (0.06)
naddtime_prop	How many times a participant selected a specific time of day for a time-based notification divided by their app use sessions.	0.03 (0.09)
nbadgescompleted_prop	How many badges a participant earned for reaching milestones in their app use or cessation journey divided by their app use sessions.	0.55 (0.45)
nbadgesviewed_prop	How many times a participant viewed a badge available to earn divided by their app use sessions.	0.01 (0.03)
nbuttonsfavorited_prop	How many times a participant favorited a content page divided by their app use sessions.	0.74 (2.44)
nbuttonshared_prop	How many times a participant shared a content page divided by their app use sessions.	0.03 (0.08)
ncardsviewed_prop	How many content pages a participant viewed divided by their app use sessions.	4.61 (4.33)
nchallengesaccepted_prop	How many times participants accepted a challenge divided by their app use sessions.	0.05 (0.07)
ncompletedemas_prop	How many EMA ^b prompts a participant completed divided by their app use sessions.	0.18 (0.17)
ncravingspressed_prop	How many times a participant pressed the “I’m Craving” button divided by their app use sessions.	0.07 (0.09)
ncustomtips_location_prop	How many times a participant entered a custom notification to receive at a specific location divided by their app use sessions.	0.00 (0.02)
ncustomtips_time_prop	How many times a participant entered a custom notification to receive at a specific time of day divided by their app use sessions.	0.01 (0.02)
nexplorecontentpages_prop	How many times a participant viewed “Tips,” “FYIs” or “Inspirations” content pages divided by their app use sessions.	1.24 (1.08)
nfeelingdownpressed_prop	How many times a participant selected the “Feeling Down” button divided by their app use sessions.	0.04 (0.06)
nfeelinggreatpressed_prop	How many times a participant selected the “I’m Great” button divided by their app use sessions.	0.10 (0.14)
nlocationtags_prop	How times a participant tagged a specific location divided by their app use sessions.	0.00 (0.02)
nnotificationsreceived_prop	How many times a participant opened a scheduled notification from the app divided by their app use sessions.	0.64 (0.31)
nprogresspressed_prop	How many times a participant pressed the “Progress” button to view their progress in their cessation journey divided by their app use sessions.	0.38 (0.35)
nquitdateset_prop	How many times participants set a new quit date divided by their app use sessions.	0.14 (0.24)
nregistrations_prop	How many times a participant registered their account divided by their app use sessions.	1.18 (1.21)
nscreensviewed_prop	How many screens a participant viewed in the app divided by their app use sessions.	7.52 (3.23)
nslippedpressed_prop	How many times a participant selected the “I Slipped” button divided by their app use sessions.	0.10 (0.13)
ntimetags_prop	How many times a participant tagged a specific time divided by their app use sessions.	0.00 (0.01)
ntotalgames_prop	How many times a participant played a game divided by their number of app use sessions.	0.10 (0.19)
Binary app feature use variables (n=2), n (%)		
noquitdate_bin	Did a participant opt not to select a quit date while setting up their profile?	23 (17.3)
quitdatereset_bin	Did a participant reset their quit date at least once?	47 (35.3)
Other variables (n=3)		
nunique_sessions, mean (SD)	A participant’s total number of app use sessions, defined as any series of actions within the app with no more than 2 minutes between actions.	54.58 (67.58)
Phonetype, n (%)	Did a participant have an iPhone?	74 (55.6)

Variable name	Definition	Values
Studyarm, n (%)	Was the participant assigned to the incentivized EMA or the nonincentivized EMA arm?	62 (46.6)

^aSML: supervised machine learning.

^bEMA: ecological momentary assessment.

The importance metrics for all 28 features in the model are displayed in Figure S2 in [Multimedia Appendix 1](#). Partial dependence plots for the 10 most important features are shown in [Figure 1](#). These plots depict the marginal effect of each feature on the probability of smoking cessation. The feature in the model with the highest variable importance was `nslippedpressed_prop`, the number of times a participant pressed the “I slipped” button divided by their number of app use sessions. By pressing this button, users access targeted content and guidance intended to help them after they had “slipped up” and smoked a cigarette. As can be seen in [Figure 1](#), this feature was negatively related to the probability of cessation, indicating that users who reported “slipping up” more often, proportional to their app use, were less likely to successfully quit smoking. The second and third most important features in the model, respectively, were `nexplorecontentpages_prop` and `nbadgescompleted_prop`. The former variable represents the number of times a participant viewed “Tips,” “FYIs,” or “Inspirations” content pages in the app divided by their number of app use sessions. The latter represents the number of badges a participant earned for reaching milestones in their cessation journey or use of the quitSTART app (eg, checking the app 5 times in 1 day) divided by their number of app use sessions. Both of these variables were positively related to cessation, showing that participants who used these app features more often were more likely to successfully quit smoking. Other features that were among the top 10 with the highest variable importance were `naddlocation_prop`, `ncompletedemas_prop`, `studyarm`, `nprogresspressed_prop`, `noquitdate_bin`, and `nfeelinggreatpressed_prop`.

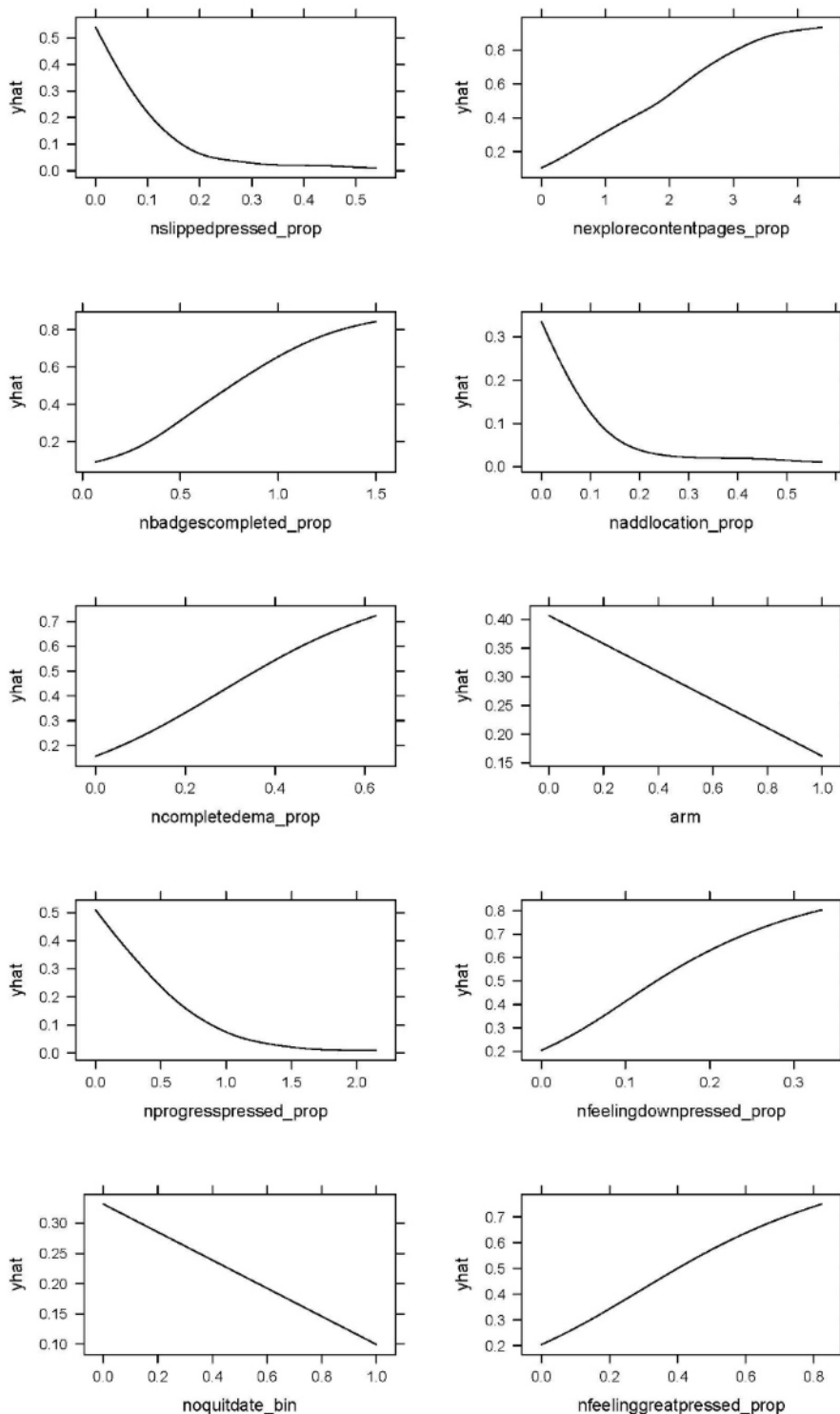
After building our SML model and assessing feature importance in the training set, we fit the model in the test set. The model’s accuracy was 0.67, and both its sensitivity and specificity were also 0.67. We retained the SML model–predicted probabilities of cessation as a variable in the test set.

Results from the 2 logistic regression models performed in the test set are summarized in Table S1 in [Multimedia Appendix 1](#). The likelihood ratio test comparing model 1, which included our set of participant characteristics that are known to be related to cessation, to a null model was not statistically significant at the $\alpha=.05$ level ($\chi^2_9=5.0$; $P=.84$). Likewise, the likelihood ratio test comparing model 2, which included all variables included in model 1, as well as the predicted probabilities of cessation from the SML model to a null model was not statistically significant ($\chi^2_{10}=7.0$; $P=.73$). The likelihood ratio test comparing model 2 to model 1 was not statistically significant ($\chi^2_1=2.0$; $P=.16$), indicating that model 2 provided a statistically equivalent fit to the data to model 1.

The variables considered for inclusion in our sensitivity analysis SML model are summarized in Table S2 in [Multimedia Appendix 1](#). Recursive feature elimination showed that the optimal number of features to include in the model was 28. The model’s accuracy in the training set was 0.88, its sensitivity was 0.94, and its specificity was 0.71. The most important feature of the model was `studyarm`, which represented the participants’ assigned study arm. The importance metrics for each feature included in the model are displayed in Figure S3 in [Multimedia Appendix 1](#) and each feature is defined in Table S2 in [Multimedia Appendix 1](#). The model’s accuracy in the test set was 0.64. Its sensitivity was 0.75 while its specificity was 0.33.

We fit 2 logistic regression models in the test set (see Table S3 in [Multimedia Appendix 1](#)) and ran a likelihood ratio test comparing the 2 models. The likelihood ratio test was not statistically significant ($\chi^2_1=0.6$; $P=.46$), indicating that model 2, which included the predicted probabilities from the SML model using continuous app feature use variables, did not provide a significantly better fit to the data than model 1.

Figure 1. Partial dependence plots depicting the predicted marginal effects on the probability of cessation for the 10 app use variables assigned the highest feature importance. The x-axis in each figure is constrained to show only values of each variable that were observed in the training set used to build the supervised machine learning model.



Discussion

Principal Findings

We developed and tested a novel approach to using SML to examine whether and how the use of specific features within a smoking cessation app predicts short-term cessation. We applied SML models to data from the quitSTART EMA Incentivization Trial to identify patterns of app feature use that predict

short-term smoking cessation. Our analysis of variable importance within this model indicated that the 3 app feature use variables that were most important for predicting cessation were the number of times participants pressed the “I Slipped” button, the number of times they viewed the “Tips,” “FYIs,” or “Inspirations” content pages, and the number of badges they completed (each expressed as a proportion of total app use sessions). We then used a likelihood ratio test comparing 2 logistic regression models to assess whether including patterns

of app feature use in our models allowed us to better predict cessation. The results of this likelihood ratio test showed that the logistic regression model that included both the SML-predicted probabilities of cessation based on participants' app feature use, as well as a set of variables reflecting participants' baseline tobacco use and demographic and personal characteristics did not fit the data better than a model that included only the latter variables. This means the accuracy of our model predicting whether participants quit smoking was not improved by including the SML-predicted probabilities. However, because only observations from the held-aside test set ($n=30$) were included in this analysis, the small n likely contributed to this null result.

This study adds to the small but growing body of literature that has gone beyond looking at the overall relationship between smoking cessation app use and smoking cessation to examine which specific app features are associated with cessation [20,21]. Some of our findings align with those from prior research. For example, our finding that completing badges is an important variable for predicting smoking cessation aligns with the finding reported by Rajani et al [20] that participants' frequency of use of gamification features, including earning badges, was associated with motivation to quit. However, there is a need for more research investigating different app features within smoking cessation apps to help maximize the potential public health impacts of smoking cessation apps. The methodological approach developed in this study could be used to guide additional research evaluating smoking cessation apps and to improve the design and refinement of such apps. While this study focused on smoking cessation, this approach could also be applied in research on apps focused on other health behaviors.

Our methodological approach could help guide further research in several ways. For example, our finding that patterns of app feature use did not predict unique variance in cessation might lead researchers to explore whether there is variability in the extent to which different groups of app users are helped by different app features. Alternatively, finding that patterns of app feature use did predict unique variance in cessation might inspire additional research investigating users' perceptions of, satisfaction with, and reasons for using the app feature use variables that were found to be important for predicting cessation.

Additionally, if an app feature uses a variable that was expected to be effective based on theory and prior research was not found to be important in predicting cessation, researchers might investigate why this was the case, considering possible explanatory factors such as design and usability issues [15,18]. This research could also inform the design of new apps, as well as the refinement of existing apps. Apps could be streamlined to only include features found to be important for cessation, which could in turn improve their cost-efficiency for app developers and usability for app users.

While this was a retrospective analysis conducted after participants had finished using quitSTART, SML models could also be applied in real time to identify current app users whose patterns of app feature use suggest they may be unlikely to quit smoking. These individuals could then be sent tailored messages

through the app to nudge them to alter their patterns of app use or connect them with additional support. For example, in this study, we found that individuals who pressed the "I slipped" button more frequently, proportional to their overall app use, were less likely to report short-term smoking cessation. If this relationship was observed in a context in which real-time intervention was possible, individuals who pressed the "I slipped" button could automatically be connected to another source of support, such as a smoking cessation counselor.

Study Limitations

Given that this was a secondary data analysis involving a relatively small convenience sample of individuals who participated in an experiment, findings from this study were not expected to be generalizable to the general population of people who smoke. Findings were also not expected to be generalizable to all quitSTART users because the experimental protocol itself may have affected some participants' app feature use. Specifically, participants in the incentivized EMA arm received compensation based on their completion of EMAs and, as a result, used that app feature more frequently than did participants in the nonincentivized EMA arm (unpublished data, 2021). The small sample size, as well as the relative rarity of our cessation outcome (about 28% of participants reported 7-day point-prevalence abstinence at 4 weeks), may also have impacted the accuracy of the SML model we fit, contributing to its suboptimal accuracy, specificity, and sensitivity in the test set. These factors may also have affected the results of our aim 2 statistical analyses.

Additionally, the app feature use variables we included in our SML model only captured the number of times a participant used a given app feature as a proportion of their overall app use or whether the participant had used an app feature at all. Future research should examine factors such as the time of day during which a participant used a given app feature or the responses given to interactive app features to get a more detailed view of the relationship between app feature use and cessation. Finally, although we accounted for several variables that might be related to cessation in our logistic regression models, the list of variables we included was not exhaustive.

Conclusions

Smartphone apps could expand the availability and use of evidence-based smoking cessation interventions, potentially helping more people quit smoking. However, there is a need for more research evaluating the effectiveness of smoking cessation apps and investigating how individuals' use of different app features impacts their likelihood of cessation. In this study, we developed and tested a novel methodological paradigm using SML to test patterns of app feature use that are most predictive of short-term smoking cessation and assess whether patterns of app feature use explain variance in cessation that is not explained by other relevant variables. We identified important app feature use variables for predicting cessation. We did not find evidence that patterns of app feature use explained variance in cessation beyond what was explained by participants' tobacco use and demographic and personal characteristics, although the small sample size likely contributed to this result. Nonetheless, the methodological approach

developed in this study could be used in future research focused broadly to inform the design and refinement of such apps. on smoking cessation apps and health behavior apps more

Acknowledgments

This work was funded by a National Cancer Institute (NCI) Trans-Fellowship Research Award and supported internally by the NCI.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Full results from logistic regression models; results and description of sensitivity analyses; and details about participant recruitment, sample allocation, and data collection.

[[DOCX File, 93 KB - ai_v3i1e51756_app1.docx](#)]

References

1. Smoking cessation: a report of the surgeon general. Centers for Disease Control and Prevention. Atlanta, GA; 2020. URL: <https://www.cdc.gov/tobacco/sgr/2020-smoking-cessation/index.html> [accessed 2024-04-11]
2. Babb S, Malarcher A, Schauer G, Asman K, Jamal A. Quitting smoking among adults—United States, 2000–2015. *MMWR Morb Mortal Wkly Rep* 2017;65(52):1457–1464 [FREE Full text] [doi: [10.15585/mmwr.mm6552a1](https://doi.org/10.15585/mmwr.mm6552a1)] [Medline: [28056007](https://pubmed.ncbi.nlm.nih.gov/28056007/)]
3. Vilardaga R, Casellas-Pujol E, McClernon JF, Garrison KA. Mobile applications for the treatment of tobacco use and dependence. *Curr Addict Rep* 2019;6(2):86–97 [FREE Full text] [doi: [10.1007/s40429-019-00248-0](https://doi.org/10.1007/s40429-019-00248-0)] [Medline: [32010548](https://pubmed.ncbi.nlm.nih.gov/32010548/)]
4. Mobile fact sheet. Pew Research Center. 2021. URL: <https://www.pewresearch.org/internet/fact-sheet/mobile/> [accessed 2022-09-06]
5. Lee J, Dallery J, Laracuenta A, Ibe I, Joseph S, Huo J, et al. A content analysis of free smoking cessation mobile applications in the USA. *J Smok Cessat* 2019;14(4):195–202. [doi: [10.1017/jsc.2019.6](https://doi.org/10.1017/jsc.2019.6)]
6. Bricker JB, Mull KE, Santiago-Torres M, Miao Z, Perski O, Di C. Smoking cessation smartphone app use over time: predicting 12-month cessation outcomes in a 2-arm randomized trial. *J Med Internet Res* 2022;24(8):e39208 [FREE Full text] [doi: [10.2196/39208](https://doi.org/10.2196/39208)] [Medline: [35831180](https://pubmed.ncbi.nlm.nih.gov/35831180/)]
7. Regmi D, Tobutt C, Shaban S. Quality and use of free smoking cessation apps for smartphones. *Int J Technol Assess Health Care* 2018;34(5):476–480. [doi: [10.1017/S0266462318000521](https://doi.org/10.1017/S0266462318000521)] [Medline: [30226123](https://pubmed.ncbi.nlm.nih.gov/30226123/)]
8. Haskins BL, Lesperance D, Gibbons P, Boudreaux ED. A systematic review of smartphone applications for smoking cessation. *Transl Behav Med* 2017;7(2):292–299 [FREE Full text] [doi: [10.1007/s13142-017-0492-2](https://doi.org/10.1007/s13142-017-0492-2)] [Medline: [28527027](https://pubmed.ncbi.nlm.nih.gov/28527027/)]
9. Bricker JB, Watson NL, Mull KE, Sullivan BM, Heffner JL. Efficacy of smartphone applications for smoking cessation: a randomized clinical trial. *JAMA Intern Med* 2020;180(11):1472–1480 [FREE Full text] [doi: [10.1001/jamainternmed.2020.4055](https://doi.org/10.1001/jamainternmed.2020.4055)] [Medline: [32955554](https://pubmed.ncbi.nlm.nih.gov/32955554/)]
10. BinDhim NF, McGeechan K, Trevena L. Smartphone Smoking Cessation Application (SSC App) trial: a multicountry double-blind automated randomised controlled trial of a smoking cessation decision-aid 'app'. *BMJ Open* 2018;8(1):e017105 [FREE Full text] [doi: [10.1136/bmjopen-2017-017105](https://doi.org/10.1136/bmjopen-2017-017105)] [Medline: [29358418](https://pubmed.ncbi.nlm.nih.gov/29358418/)]
11. Bricker JB, Mull KE, Kientz JA, Vilardaga R, Mercer LD, Akioka KJ, et al. Randomized, controlled pilot trial of a smartphone app for smoking cessation using acceptance and commitment therapy. *Drug Alcohol Depend* 2014;143:87–94 [FREE Full text] [doi: [10.1016/j.drugalcdep.2014.07.006](https://doi.org/10.1016/j.drugalcdep.2014.07.006)] [Medline: [25085225](https://pubmed.ncbi.nlm.nih.gov/25085225/)]
12. Browne J, Halverson TF, Vilardaga R. Engagement with a digital therapeutic for smoking cessation designed for persons with psychiatric illness fully mediates smoking outcomes in a pilot randomized controlled trial. *Transl Behav Med* 2021;11(9):1717–1725. [doi: [10.1093/tbm/ibab100](https://doi.org/10.1093/tbm/ibab100)] [Medline: [34347865](https://pubmed.ncbi.nlm.nih.gov/34347865/)]
13. Hoepfner BB, Hoepfner SS, Seaboyer L, Schick MR, Wu GWY, Bergman BG, et al. How smart are smartphone apps for smoking cessation? A content analysis. *Nicotine Tob Res* 2016;18(5):1025–1031 [FREE Full text] [doi: [10.1093/ntr/ntv117](https://doi.org/10.1093/ntr/ntv117)] [Medline: [26045249](https://pubmed.ncbi.nlm.nih.gov/26045249/)]
14. Barroso-Hurtado M, Suárez-Castro D, Martínez-Vispo C, Becoña E, López-Durán A. Smoking cessation apps: a systematic review of format, outcomes, and features. *Int J Environ Res Public Health* 2021;18(21):11664 [FREE Full text] [doi: [10.3390/ijerph182111664](https://doi.org/10.3390/ijerph182111664)] [Medline: [34770178](https://pubmed.ncbi.nlm.nih.gov/34770178/)]
15. Paige SR, Alber JM, Stellefson ML, Krieger JL. Missing the mark for patient engagement: mHealth literacy strategies and behavior change processes in smoking cessation apps. *Patient Educ Couns* 2018;101(5):951–955 [FREE Full text] [doi: [10.1016/j.pec.2017.11.006](https://doi.org/10.1016/j.pec.2017.11.006)] [Medline: [29153592](https://pubmed.ncbi.nlm.nih.gov/29153592/)]

16. Rajani NB, Weth D, Mastellos N, Filippidis FT. Adherence of popular smoking cessation mobile applications to evidence-based guidelines. *BMC Public Health* 2019;19(1):743 [FREE Full text] [doi: [10.1186/s12889-019-7084-7](https://doi.org/10.1186/s12889-019-7084-7)] [Medline: [31196062](https://pubmed.ncbi.nlm.nih.gov/31196062/)]
17. Ubhi HK, Michie S, Kotz D, van Schayck OCP, Selladurai A, West R. Characterising smoking cessation smartphone applications in terms of behaviour change techniques, engagement and ease-of-use features. *Transl Behav Med* 2016;6(3):410-417 [FREE Full text] [doi: [10.1007/s13142-015-0352-x](https://doi.org/10.1007/s13142-015-0352-x)] [Medline: [27528530](https://pubmed.ncbi.nlm.nih.gov/27528530/)]
18. Bendotti H, Lawler S, Ireland D, Gartner C, Hides L, Marshall HM. What do people want in a smoking cessation app? An analysis of user reviews and app quality. *Nicotine Tob Res* 2022;24(2):169-177. [doi: [10.1093/ntr/ntab174](https://doi.org/10.1093/ntr/ntab174)] [Medline: [34460922](https://pubmed.ncbi.nlm.nih.gov/34460922/)]
19. Budenz A, Wiseman KP, Keefe B, Prutzman Y. User engagement with mood-related content on the National Cancer Institute Smokefree.Gov initiative cessation resources. *Health Educ Behav* 2022;49(4):613-617 [FREE Full text] [doi: [10.1177/10901981211073736](https://doi.org/10.1177/10901981211073736)] [Medline: [35112581](https://pubmed.ncbi.nlm.nih.gov/35112581/)]
20. Rajani NB, Mastellos N, Filippidis FT. Impact of gamification on the self-efficacy and motivation to quit of smokers: observational study of two gamified smoking cessation mobile apps. *JMIR Serious Games* 2021;9(2):e27290 [FREE Full text] [doi: [10.2196/27290](https://doi.org/10.2196/27290)] [Medline: [33904824](https://pubmed.ncbi.nlm.nih.gov/33904824/)]
21. Heffner JL, Vilardaga R, Mercer LD, Kientz JA, Bricker JB. Feature-level analysis of a novel smartphone application for smoking cessation. *Am J Drug Alcohol Abuse* 2015;41(1):68-73 [FREE Full text] [doi: [10.3109/00952990.2014.977486](https://doi.org/10.3109/00952990.2014.977486)] [Medline: [25397860](https://pubmed.ncbi.nlm.nih.gov/25397860/)]
22. Hoeppe BB, Siegel KR, Carlon HA, Kahler CW, Park ER, Taylor ST, et al. Feature-level analysis of a smoking cessation smartphone app based on a positive psychology approach: prospective observational study. *JMIR Form Res* 2022;6(7):e38234 [FREE Full text] [doi: [10.2196/38234](https://doi.org/10.2196/38234)] [Medline: [35900835](https://pubmed.ncbi.nlm.nih.gov/35900835/)]
23. Abo-Tabik M, Costen N, Darby J, Benn Y. Towards a smart smoking cessation app: a 1D-CNN model predicting smoking events. *Sensors (Basel)* 2020;20(4):1099 [FREE Full text] [doi: [10.3390/s20041099](https://doi.org/10.3390/s20041099)] [Medline: [32079359](https://pubmed.ncbi.nlm.nih.gov/32079359/)]
24. Abo-Tabik M, Benn Y, Costen N. Are machine learning methods the future for smoking cessation apps? *Sensors (Basel)* 2021;21(13):4254 [FREE Full text] [doi: [10.3390/s21134254](https://doi.org/10.3390/s21134254)] [Medline: [34206167](https://pubmed.ncbi.nlm.nih.gov/34206167/)]
25. Prutzman YM, Wiseman KP, Grady MA, Budenz A, Grenen EG, Vercammen LK, et al. Using digital technologies to reach tobacco users who want to quit: evidence from the National Cancer Institute's Smokefree.gov initiative. *Am J Prev Med* 2021;60(3 Suppl 2):S172-S184 [FREE Full text] [doi: [10.1016/j.amepre.2020.08.008](https://doi.org/10.1016/j.amepre.2020.08.008)] [Medline: [33663705](https://pubmed.ncbi.nlm.nih.gov/33663705/)]
26. Kroenke K, Spitzer RL, Williams JB. The PHQ-9: validity of a brief depression severity measure. *J Gen Intern Med* 2001;16(9):606-613 [FREE Full text] [doi: [10.1046/j.1525-1497.2001.016009606.x](https://doi.org/10.1046/j.1525-1497.2001.016009606.x)] [Medline: [11556941](https://pubmed.ncbi.nlm.nih.gov/11556941/)]
27. Heatherton TF, Kozlowski LT, Frecker RC, Fagerström KO. The Fagerström test for nicotine dependence: a revision of the Fagerström Tolerance Questionnaire. *Br J Addict* 1991;86(9):1119-1127. [doi: [10.1111/j.1360-0443.1991.tb01879.x](https://doi.org/10.1111/j.1360-0443.1991.tb01879.x)] [Medline: [1932883](https://pubmed.ncbi.nlm.nih.gov/1932883/)]
28. Dinga R, Schmaal L, Penninx BWJH, Veltman DJ, Marquand AF. Controlling for effects of confounding variables on machine learning predictions. *bioRxiv* 2020 Preprint posted online August 18, 2020. [doi: [10.1101/2020.08.17.255034](https://doi.org/10.1101/2020.08.17.255034)]
29. Greenwell BM. pdp: an R package for constructing partial dependence plots. *R J* 2017;9(1):421-436 [FREE Full text] [doi: [10.32614/rj-2017-016](https://doi.org/10.32614/rj-2017-016)]
30. Storr CL, Reboussin BA, Anthony JC. The Fagerström test for nicotine dependence: a comparison of standard scoring and latent class analysis approaches. *Drug Alcohol Depend* 2005;80(2):241-250. [doi: [10.1016/j.drugalcdep.2004.04.021](https://doi.org/10.1016/j.drugalcdep.2004.04.021)] [Medline: [15908142](https://pubmed.ncbi.nlm.nih.gov/15908142/)]

Abbreviations

- EMA:** ecological momentary assessment
PHQ-9: Patient Health Questionnaire-9
RCT: randomized controlled trial
SML: supervised machine learning

Edited by Z Yin; submitted 15.08.23; peer-reviewed by T Dang, A Kundu, N Fradkin; comments to author 05.11.23; revised version received 29.02.24; accepted 04.03.24; published 22.05.24.

Please cite as:

Siegel LN, Wiseman KP, Budenz A, Prutzman Y

Identifying Patterns of Smoking Cessation App Feature Use That Predict Successful Quitting: Secondary Analysis of Experimental Data Leveraging Machine Learning

JMIR AI 2024;3:e51756

URL: <https://ai.jmir.org/2024/1/e51756>

doi: [10.2196/51756](https://doi.org/10.2196/51756)

PMID: [38875564](https://pubmed.ncbi.nlm.nih.gov/38875564/)

©Leeann Nicole Siegel, Kara P Wiseman, Alex Budenz, Yvonne Prutzman. Originally published in JMIR AI (<https://ai.jmir.org>), 22.05.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR AI, is properly cited. The complete bibliographic information, a link to the original publication on <https://www.ai.jmir.org/>, as well as this copyright and license information must be included.

Original Paper

Use of Deep Neural Networks to Predict Obesity With Short Audio Recordings: Development and Usability Study

Jingyi Huang¹, MA; Peiqi Guo², MISM, MEM; Sheng Zhang³, PhD; Mengmeng Ji⁴, MBBS, PhD; Ruopeng An^{2,5}, MPP, PhD

¹School of Economics and Management, Shanghai University of Sport, Shanghai, China

²Brown School, Washington University in St. Louis, St. Louis, MO, United States

³School of Journalism and Communication, Shanghai University of Sport, Shanghai, China

⁴Division of Public Health Sciences, Department of Surgery, Washington University School of Medicine in St. Louis, St. Louis, MO, United States

⁵Division of Data and Computational Sciences, Washington University in St. Louis, St. Louis, MO, United States

Corresponding Author:

Sheng Zhang, PhD

School of Journalism and Communication

Shanghai University of Sport

650 Hengren Road

Yangpu District

Shanghai, 200000

China

Phone: 86 18017355353

Email: zhsheng1@126.com

Abstract

Background: The escalating global prevalence of obesity has necessitated the exploration of novel diagnostic approaches. Recent scientific inquiries have indicated potential alterations in voice characteristics associated with obesity, suggesting the feasibility of using voice as a noninvasive biomarker for obesity detection.

Objective: This study aims to use deep neural networks to predict obesity status through the analysis of short audio recordings, investigating the relationship between vocal characteristics and obesity.

Methods: A pilot study was conducted with 696 participants, using self-reported BMI to classify individuals into obesity and nonobesity groups. Audio recordings of participants reading a short script were transformed into spectrograms and analyzed using an adapted YOLOv8 model (Ultralytics). The model performance was evaluated using accuracy, recall, precision, and F_1 -scores.

Results: The adapted YOLOv8 model demonstrated a global accuracy of 0.70 and a macro F_1 -score of 0.65. It was more effective in identifying nonobesity (F_1 -score of 0.77) than obesity (F_1 -score of 0.53). This moderate level of accuracy highlights the potential and challenges in using vocal biomarkers for obesity detection.

Conclusions: While the study shows promise in the field of voice-based medical diagnostics for obesity, it faces limitations such as reliance on self-reported BMI data and a small, homogenous sample size. These factors, coupled with variability in recording quality, necessitate further research with more robust methodologies and diverse samples to enhance the validity of this novel approach. The findings lay a foundational step for future investigations in using voice as a noninvasive biomarker for obesity detection.

(JMIR AI 2024;3:e54885) doi:[10.2196/54885](https://doi.org/10.2196/54885)

KEYWORDS

obesity; obese; overweight; voice; vocal; vocal cord; vocal cords; voice-based; machine learning; ML; artificial intelligence; AI; algorithm; algorithms; predictive model; predictive models; predictive analytics; predictive system; practical model; practical models; early warning; early detection; deep neural network; deep neural networks; DNN; artificial neural network; artificial neural networks; deep learning

Introduction

Obesity has emerged as a prominent global health concern, with its prevalence nearly tripling since 1975 and affecting a significant portion of the population worldwide [1]. This increase is especially pronounced in developing nations, partially owing to shifts in lifestyle and dietary habits [2]. Obesity serves as a precursor to various medical conditions including, but not limited to, type 2 diabetes, cardiovascular diseases, certain forms of cancer, and musculoskeletal disorders, significantly contributing to the global disease burden and elevating premature mortality rates [3]. The increased health care expenditures and reduced productivity adversely impacted the regional economy [4].

While the broad ramifications of obesity are well documented, recent scientific inquiries have begun to elucidate the potential alterations in voice characteristics that may be concurrent with obesity [5,6]. Several mechanisms are postulated to explain these alterations in vocal attributes. The deposition of adipose tissue near the vocal folds and larynx may influence vocal resonance and pitch, often resulting in variations in voice quality [7]. Restrictive lung patterns associated with obesity may lead to compromised lung volumes and capacities, subsequently impacting subglottal pressures essential for phonation [8]. Obesity induces a chronic inflammatory state, potentially altering the composition and viscosity of vocal fold tissues and affecting parameters such as jitter and shimmer [9]. In addition, the hormonal imbalances often seen in obesity can impact the elasticity and tension of laryngeal tissues, thereby influencing voice characteristics [10].

Given these insights, voice-based markers have emerged as a pioneering approach to assessing obesity [11]. The prospect of using noninvasive and readily accessible audio recordings may pave the way for advancements in diagnostic methodologies, overcoming the constraints inherent to current obesity assessment techniques [12]. This innovative method holds the potential to inform preventive health care strategies by enabling the extraction of critical health information from voice, allowing for the development of scalable, real-time, and accurate health-monitoring systems. The implications of such advancements could be especially significant in regions with limited resources, facilitating early interventions and alleviating the compounded health and economic repercussions associated with obesity. Delving into the intricate relationship between voice characteristics and obesity may enhance our understanding and propel the evolution of novel diagnostic and monitoring tools, presenting opportunities for refined strategies in obesity management and prevention.

Artificial intelligence (AI), characterized by machine and deep learning techniques, has become increasingly popular in exploring and addressing the multifaceted challenges associated with obesity [13,14]. For instance, studies have used deep neural

network models to analyze face portrait photographs to predict obesity status and the risk of diabetes, showcasing the versatility and efficacy of AI in medical diagnoses and risk assessments [15]. These applications exemplify the transformative potential of AI in deriving insightful correlations and predictive analytics in the context of obesity, allowing for the development of sophisticated and nuanced approaches to studying and managing this prevalent condition.

This pilot study pioneers the exploration of using deep neural network models to predict individuals' obesity status through analyses of short audio recordings. Participants were recorded while reading a prewritten script, and the models were developed to discern potential associations between vocal characteristics and obesity. This study constitutes the initial endeavor to explore the relationship between obesity and voice, highlighting an uncharted intersection in obesity research. Although preliminary, the study lays the groundwork in this novel domain, and relevant findings may inspire future research in voice-related health diagnostics.

Methods

Data

We conducted a standardized web-based survey to gather demographic information (gender and age), self-reported anthropometric measurements (height and weight), disease histories, and brief audio recordings from participants (see [Multimedia Appendices 1 and 2](#)). The participants were instructed to read a short Mandarin paragraph provided in the survey and record it using their mobile phones. Consequently, the final analysis comprised 696 participants, including 500 females and 196 males, with an average age of 24 years.

We classified study participants into 2 groups, obesity (271/696, 38.9%) and nonobesity (425/696, 61.1%), based on the standard BMI threshold of $\geq 28 \text{ kg/m}^2$ in the Chinese population [16].

A spectrogram is a visual representation of the spectrum of frequencies in a sound signal as they vary with time, serving as an essential tool for feature extraction in audio classification tasks. Audio recordings were standardized to the WAV format and then transformed into spectrograms. The preprocessed data set was randomly partitioned into a training set of 591 audio files (591/696, 85%) and a test set of 105 files (105/696, 15%).

Data augmentation on spectrograms involves applying various techniques such as time stretching, noise injection, and frequency masking to enhance the diversity and robustness of the data set, thereby improving the performance of machine learning models in audio classification. Data augmentation was used to balance the training set, ensuring equal representations of images labeled as obesity and nonobesity. Subsequently, a 5-fold cross-validation was performed on the balanced training set. Our workflow is illustrated in [Figure 1](#).

Figure 1. Research workflow.



Ethical Considerations

The study was approved by the Shanghai University of Sport Ethics Committee (institutional review board #102772022RT065), with written informed consent obtained from each study participant. After negotiations, each participant received 10 yuan as compensation for participating in the study, and the data of each participant were anonymized.

Model

We developed a neural network model to predict an individual's obesity status using spectrogram data. Adapting the YOLO (You Only Look Once) framework [17], which is renowned for real-time object detection and image segmentation in computer vision, we fine-tuned the pretrained YOLOv8 model for our voice-based obesity classifier. To enhance model performance,

we used techniques such as batch normalization, learning rate optimization, label smoothing, and early stopping. This model was constructed using Python (version 3.10.12; Python Software Foundation) and was accelerated using a Tesla V100 GPU (NVIDIA).

A comparison of the predictive performances of corresponding models applying 2 main feature extraction approaches in speech recognition was conducted. One is based on signal parameter extraction, such as Mel-frequency cepstral coefficients and Mel-filter bank features, while the other is based on spectrogram images. Table 1 delineates the performance metrics of multiple models across varied feature sets. The YOLOv8 model we applied exhibited higher performance, which is specified in italics.

Table 1. Overall performances of various models.

Features and model	F_1 -score	Sensitivity	PPV ^a	Accuracy
Spectrogram				
Yolov8	<i>0.65^b</i>	<i>0.69</i>	<i>0.65</i>	<i>0.70</i>
CNN ^c	0.59	0.58	0.61	0.60
MFCCs^d+Delta-Delta				
CNN	0.57	0.56	0.58	0.62
RandomForest	0.58	0.56	0.59	0.63
MLP ^e	0.56	0.57	0.56	0.56
MFCCs+Mel^f				
CNN	0.59	0.57	0.61	0.64
RandomForest	0.58	0.57	0.60	0.63
MLP	0.55	0.55	0.55	0.57

^aPPV: positive predictive value.

^bItalics indicates higher performance.

^cCNN: convolutional neural network.

^dMFCC: Mel-frequency cepstral coefficient.

^eMLP: multilayer perceptron.

^fMel: Mel-filter bank features.

Results

Figure 2 shows 2 example spectrogram images transformed from audio files labeled as nonobesity and obesity. In terms of the spectrogram, horizontal axes indicate time in milliseconds. Vertical axes indicate the frequency in hertz (Hz). Brightness indicates decibel level; the brighter it is, the higher the decibel level. The stripes in the spectrogram reflect the fundamental characteristics of a speaker's voice. Bars that are relatively parallel to the horizontal axis correspond to the formant. The distance between dark stripes perpendicular to the horizontal axis represents the period of fundamental frequency. Formant and fundamental periods are closely related to the state of the vocal tract structures.

Figure 3 depicts the 5-fold cross-validation training process. The training loss gradually declined from around 0.15 to near zero by epoch 80. During epochs 0-80, the validation loss primarily decreased but with some fluctuations. From epochs 60-150, it began to stabilize around 0.25, with no substantial reductions afterward. The peak model performance was achieved at epoch 120, with a validation loss of 0.26 and an associated training loss of 0.10. Trail 4 displayed different epoch numbers due to a relatively small sample size and training fluctuations, which triggered the early-stop feature of the YOLOv8 model. During the training process, the curves of train loss and validation loss did not perfectly coincide at the end. However, the consistent downward and convergent trend of both indicated that the model was trained normally without overfitting or underfitting.

Figure 2. Spectrogram images labeled nonobesity (left) and obesity (right).

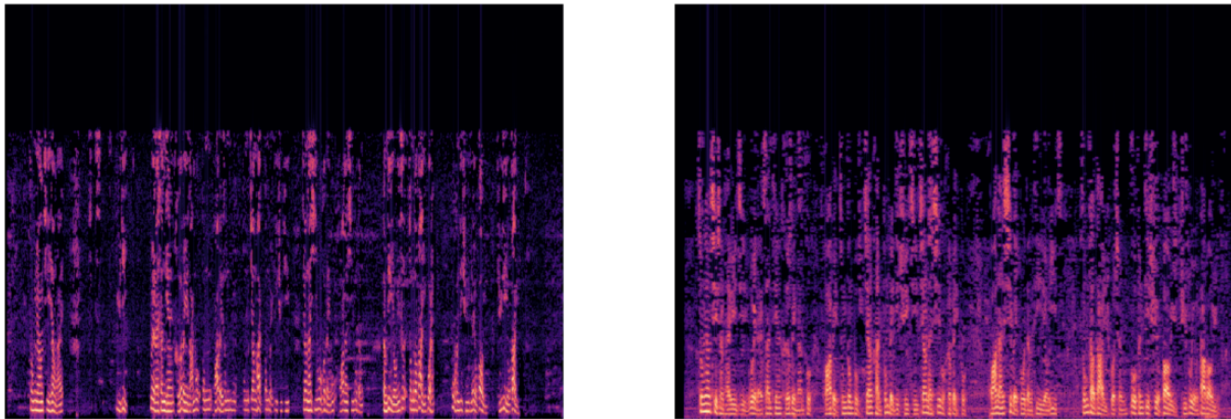


Figure 3. Model training using 5-fold cross-validation. Train loss: training loss; val loss: validation loss.

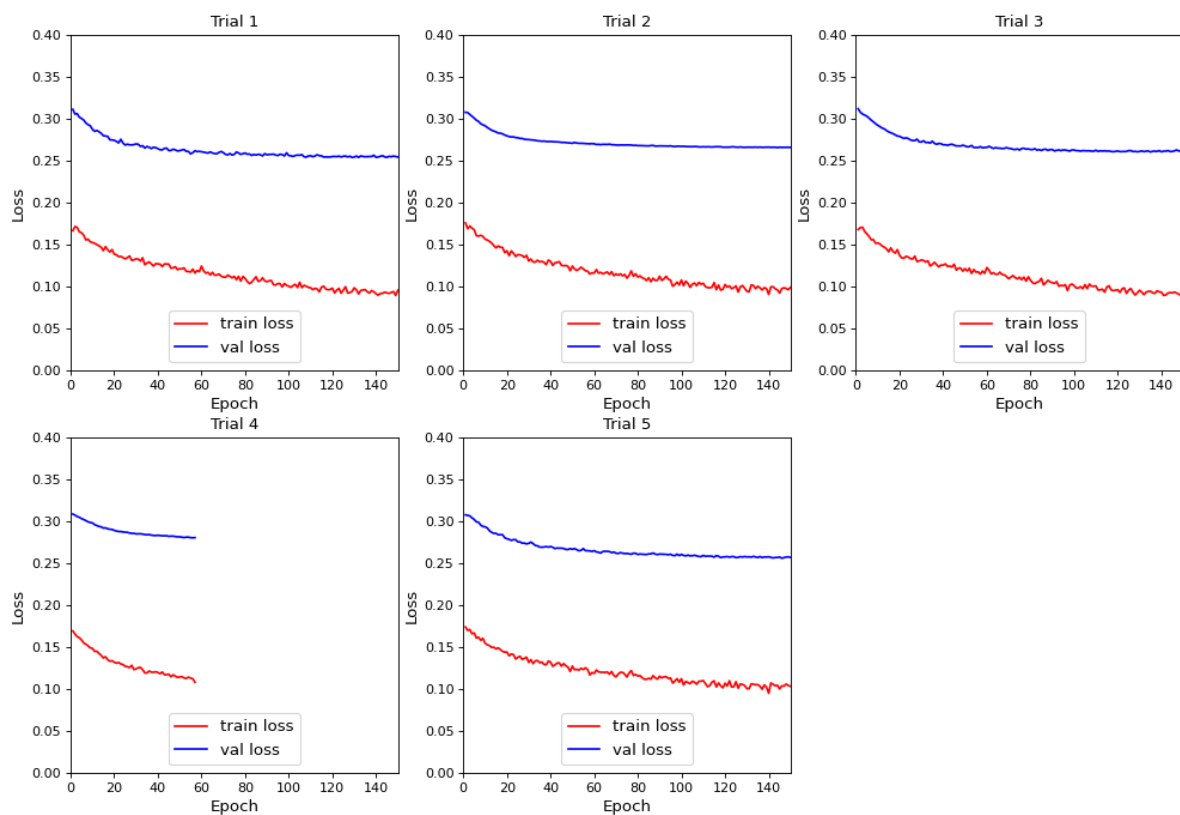


Table 2 reports the model performance on the test set. For the obesity category, the model yielded an F_1 -score of 0.53, with a recall (sensitivity) of 0.67 and a precision (positive predictive value) of 0.44. The model achieved an F_1 -score of 0.77 for

nonobesity classifications, with a recall of 0.70 and a precision of 0.86. The overall model performance across both categories was characterized by a macro F_1 -score of 0.65, a recall of 0.69, a precision of 0.65, and a global accuracy of 0.70.

Table 2. YOLOv8 model performance on the test set.

	F_1 -score	Sensitivity	PPV ^a	Accuracy
Obesity	0.53	0.67	0.44	— ^b
Nonobesity	0.77	0.70	0.86	—
Overall	0.65	0.69	0.65	0.70

^aPPV: positive predictive value.

^bNot available.

Discussion

This study explored the use of deep neural networks, specifically an adapted YOLOv8 model, to predict obesity status from short audio recordings. This approach aimed to identify potential relationships between vocal characteristics and obesity. Our results indicate a moderate level of accuracy in the model performance, with a global accuracy of 0.70 and a macro F_1 -score of 0.65. The model demonstrated a higher effectiveness in identifying nonobesity cases, as reflected by an F_1 -score of 0.77, compared with a lower F_1 -score of 0.53 for obesity classifications. These outcomes suggest that while the model shows promise, there is a need for further refinement to enhance its precision and reliability in obesity detection using vocal biomarkers.

In the context of medical diagnostics, the use of voice as a biomarker has been an emerging area of interest [18], although its application in obesity identification remains relatively unexplored. Historically, voice analysis has been successfully used in the detection of various health conditions, such as Parkinson disease, where vocal cord and speech pattern changes are indicative of the disease's progression [19]. Similarly, in respiratory diseases, voice alterations often reflect changes in lung function and airflow [20]. The rationale behind these applications is that physiological changes, whether due to neurological, respiratory, or other systemic conditions, can manifest in measurable changes in voice characteristics [21].

The aim of our study to correlate voice characteristics with obesity aligns with this emerging trend but ventures into a relatively uncharted domain. Obesity, being a complex condition with multifactorial etiologies, may not exhibit as direct a relationship with vocal changes as seen in neurological or respiratory illnesses [22]. Nonetheless, the premise that obesity can induce physiological alterations, such as in the laryngeal tissues and respiratory system [23], provides a theoretical foundation for our exploration. The moderate success of our model in distinguishing obesity from nonobesity cases indicates a potential, albeit complex, link between obesity and voice characteristics.

The findings of this study contribute to the expanding literature on noninvasive diagnostic methods. Traditional obesity diagnosis primarily relies on physical measurements such as BMI and waist circumference, which have their limitations, including the inability to assess body fat distribution and differentiate between fat and muscle mass [24]. The prospect of supplementing these methods with voice analysis could offer a more holistic and convenient approach to obesity assessment.

Using deep neural networks, short audio recordings can predict obesity status, offering practical applications in preventive medicine, telemedicine, and public health research. It enables noninvasive early screening for obesity and related health issues such as obstructive sleep apnea [25], provides objective measures in telemedicine, and offers a cost-effective data collection approach for obesity prevalence research.

However, our study's moderate accuracy underscores the challenges inherent in this novel diagnostic pathway. It highlights the need for further research to better understand the nuances of how obesity might specifically alter vocal characteristics and how these changes can be more accurately captured and interpreted by advanced neural network models.

This study faces several key limitations. Foremost, the reliance on self-reported BMI introduces potential inaccuracies due to measurement errors and social desirability bias [26], compromising the model's accuracy in obesity classification. In addition, the use of a small, convenience sample limits the statistical power and generalizability of our findings, as it may not adequately represent the broader population. Variability in audio recording quality, resulting from participants using their own mobile phones, further challenges the consistency of the input data. The demographic homogeneity of the sample and the lack of consideration for other factors influencing voice characteristics, such as lifestyle choices, restrict the applicability of our findings to a wider, more diverse population. These limitations collectively underscore the need for more robust methodologies and diverse participant samples in future research to enhance the validity and applicability of voice analysis in obesity detection.

Future research should prioritize conducting a longitudinal cohort study to analyze voice changes in individuals transitioning from lean to obese phases. This will deepen our understanding of voice changes during obesity progression and enable the extraction of vocal characteristic features across different stages of obesity. Ultimately, such an approach may aid in developing causal links between obesity and vocal changes.

In sum, while our study presents an innovative approach to obesity detection and adds to the growing body of research on voice-based medical diagnostics, it also emphasizes the complexity of this endeavor and the necessity for continued research and development in this area. The potential of using voice as a noninvasive biomarker for obesity is an intriguing prospect, and our findings, though moderate in their current state, lay the groundwork for future investigations to refine and enhance this novel diagnostic method.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Codebook and questionnaire (in Chinese).

[[PDF File \(Adobe PDF File\), 327 KB - ai_v3i1e54885_app1.pdf](#)]

Multimedia Appendix 2

Codebook and questionnaire (in English).

[\[DOCX File , 17 KB - ai_v3i1e54885_app2.docx \]](#)

References

1. Blüher M. Obesity: global epidemiology and pathogenesis. *Nat Rev Endocrinol* 2019;15(5):288-298. [doi: [10.1038/s41574-019-0176-8](https://doi.org/10.1038/s41574-019-0176-8)] [Medline: [30814686](https://pubmed.ncbi.nlm.nih.gov/30814686/)]
2. Popkin BM, Adair LS, Ng SW. Global nutrition transition and the pandemic of obesity in developing countries. *Nutr Rev* 2012;70(1):3-21 [FREE Full text] [doi: [10.1111/j.1753-4887.2011.00456.x](https://doi.org/10.1111/j.1753-4887.2011.00456.x)] [Medline: [22221213](https://pubmed.ncbi.nlm.nih.gov/22221213/)]
3. Pi-Sunyer X. The medical risks of obesity. *Postgrad Med* 2009;121(6):21-33 [FREE Full text] [doi: [10.3810/pgm.2009.11.2074](https://doi.org/10.3810/pgm.2009.11.2074)] [Medline: [19940414](https://pubmed.ncbi.nlm.nih.gov/19940414/)]
4. Tremmel M, Gerdtham UG, Nilsson PM, Saha S. Economic burden of obesity: a systematic literature review. *Int J Environ Res Public Health* 2017;14(4):435 [FREE Full text] [doi: [10.3390/ijerph14040435](https://doi.org/10.3390/ijerph14040435)] [Medline: [28422077](https://pubmed.ncbi.nlm.nih.gov/28422077/)]
5. Munjal S, Sharma A, Chhabra N, Panda N. Perceptual, aerodynamic and acoustic evaluation of vocal characteristics in subjects with obesity. *J Voice* 2024;38(3):660-665. [doi: [10.1016/j.jvoice.2021.10.019](https://doi.org/10.1016/j.jvoice.2021.10.019)] [Medline: [34969555](https://pubmed.ncbi.nlm.nih.gov/34969555/)]
6. Bosso JR, Martins RHG, Pessin ABB, Tavares ELM, Leite CV, Naresse LE. Vocal characteristics of patients with morbid obesity. *J Voice* 2021;35(2):329.e7-329.e11. [doi: [10.1016/j.jvoice.2019.09.012](https://doi.org/10.1016/j.jvoice.2019.09.012)] [Medline: [31648859](https://pubmed.ncbi.nlm.nih.gov/31648859/)]
7. Solomon N, Helou L, Dietrich-Burns K, Stojadinovic A. Do obesity and weight loss affect vocal function? *Semin Speech Lang* 2011;32(1):31-42. [doi: [10.1055/s-0031-1271973](https://doi.org/10.1055/s-0031-1271973)] [Medline: [21491357](https://pubmed.ncbi.nlm.nih.gov/21491357/)]
8. Zammit C, Liddicoat H, Moonsie I, Makker H. Obesity and respiratory diseases. *Int J Gen Med* 2010;3:335-343 [FREE Full text] [doi: [10.2147/IJGM.S11926](https://doi.org/10.2147/IJGM.S11926)] [Medline: [21116339](https://pubmed.ncbi.nlm.nih.gov/21116339/)]
9. Bonilha HS, White L, Kuckhahn K, Gerlach TT, Deliyski DD. Vocal fold mucus aggregation in persons with voice disorders. *J Commun Disord* 2012;45(4):304-311 [FREE Full text] [doi: [10.1016/j.jcomdis.2012.03.001](https://doi.org/10.1016/j.jcomdis.2012.03.001)] [Medline: [22510352](https://pubmed.ncbi.nlm.nih.gov/22510352/)]
10. de Souza LBR, Santos MMD. Body mass index and acoustic voice parameters: is there a relationship? *Braz J Otorhinolaryngol* 2018;84(4):410-415 [FREE Full text] [doi: [10.1016/j.bjorl.2017.04.003](https://doi.org/10.1016/j.bjorl.2017.04.003)] [Medline: [28545946](https://pubmed.ncbi.nlm.nih.gov/28545946/)]
11. da Cunha MGB, Passerotti GH, Weber R, Zilberstein B, Cecconello I. Voice feature characteristic in morbid obese population. *Obes Surg* 2011;21(3):340-344. [doi: [10.1007/s11695-009-9959-7](https://doi.org/10.1007/s11695-009-9959-7)] [Medline: [19763710](https://pubmed.ncbi.nlm.nih.gov/19763710/)]
12. Amato F, Fasani M, Raffaelli G. Obesity and gastro-esophageal reflux voice disorders: a machine learning approach. 2022 Presented at: IEEE International Symposium on Medical Measurements and Applications; June 22, 2022; Messina, Italy. [doi: [10.1109/memea54994.2022.9856574](https://doi.org/10.1109/memea54994.2022.9856574)]
13. An R, Shen J, Xiao Y. Applications of artificial intelligence to obesity research: scoping review of methodologies. *J Med Internet Res* 2022;24(12):e40589 [FREE Full text] [doi: [10.2196/40589](https://doi.org/10.2196/40589)] [Medline: [36476515](https://pubmed.ncbi.nlm.nih.gov/36476515/)]
14. Kumar Y, Koul A, Singla R, Ijaz MF. Artificial intelligence in disease diagnosis: a systematic literature review, synthesizing framework and future research agenda. *J Ambient Intell Humaniz Comput* 2023;14(7):8459-8486 [FREE Full text] [doi: [10.1007/s12652-021-03612-z](https://doi.org/10.1007/s12652-021-03612-z)] [Medline: [35039756](https://pubmed.ncbi.nlm.nih.gov/35039756/)]
15. Chanda A, Chatterjee S. Predicting obesity using facial pictures during COVID-19 pandemic. *Biomed Res Int* 2021;2021:6696357. [doi: [10.1155/2021/6696357](https://doi.org/10.1155/2021/6696357)] [Medline: [33778081](https://pubmed.ncbi.nlm.nih.gov/33778081/)]
16. Body weight determination for adults. National Health Commission of the People's Republic of China. 2013. URL: <http://www.nhc.gov.cn/wjw/yinyang/201308/a233d450fdb47c5ad4f08b7e394d1e8.shtml> [accessed 2023-08-31]
17. Redmon J, Divvala S, Girshick R, Farhadi A. You Only Look Once: unified, real-time object detection. arXiv Preprint posted online on Jun 8, 2015 [FREE Full text] [doi: [10.48550/arXiv.1506.02640](https://doi.org/10.48550/arXiv.1506.02640)]
18. Fagherazzi G, Fischer A, Ismael M, Despotovic V. Voice for health: the use of vocal biomarkers from research to clinical practice. *Digit Biomark* 2021;5(1):78-88 [FREE Full text] [doi: [10.1159/000515346](https://doi.org/10.1159/000515346)] [Medline: [34056518](https://pubmed.ncbi.nlm.nih.gov/34056518/)]
19. Amato F, Saggio G, Cesarini V, Olmo G, Costantini G. Machine learning- and statistical-based voice analysis of parkinson's disease patients: a survey. *Expert Syst Appl* 2023;219:119651. [doi: [10.1016/j.eswa.2023.119651](https://doi.org/10.1016/j.eswa.2023.119651)]
20. Dejonckere PH. Assessment of Voice and Respiratory Function. New York, NY: Springer eBooks; 2009:11-26.
21. Tong JY, Sataloff RT. Respiratory function and voice: the role for airflow measures. *J Voice* 2022;36(4):542-553. [doi: [10.1016/j.jvoice.2020.07.019](https://doi.org/10.1016/j.jvoice.2020.07.019)] [Medline: [32981809](https://pubmed.ncbi.nlm.nih.gov/32981809/)]
22. Celebi S, Yelken K, Develioglu ON, Topak M, Celik O, Ipek HD, et al. Acoustic, perceptual and aerodynamic voice evaluation in an obese population. *J Laryngol Otol* 2013;127(10):987-990. [doi: [10.1017/S0022215113001916](https://doi.org/10.1017/S0022215113001916)] [Medline: [24124897](https://pubmed.ncbi.nlm.nih.gov/24124897/)]
23. Salome CM, King GG, Berend N. Physiology of obesity and effects on lung function. *J Appl Physiol* (1985) 2010;108(1):206-211 [FREE Full text] [doi: [10.1152/jappphysiol.00694.2009](https://doi.org/10.1152/jappphysiol.00694.2009)] [Medline: [19875713](https://pubmed.ncbi.nlm.nih.gov/19875713/)]
24. Kok P, Seidell JC, Meinders AE. De waarde en de beperkingen van de 'body mass index' (BMI) voor het bepalen van het gezondheidsrisico van overgewicht en obesitas [The value and limitations of the body mass index (BMI) in the assessment of the health risks of overweight and obesity]. *Ned Tijdschr Geneesk* 2004;148(48):2379-2382. [doi: [10.47671/tvg.65.05.2000493](https://doi.org/10.47671/tvg.65.05.2000493)]
25. Bonsignore MR. Obesity and obstructive sleep apnea. *Handbook Exp Pharmacol* 2022;274:181-201. [doi: [10.1007/164_2021_558](https://doi.org/10.1007/164_2021_558)]

26. Bauhoff S. Systematic self-report bias in health data: impact on estimating cross-sectional and treatment effects. *Health Serv Outcomes Res Methodol* 2011;11(1-2):44-53. [doi: [10.1007/s10742-011-0069-3](https://doi.org/10.1007/s10742-011-0069-3)]

Abbreviations

AI: artificial intelligence

YOLO: You Only Look Once

Edited by K El Emam, B Malin; submitted 26.11.23; peer-reviewed by D Singh, SF Qadri, L Huang; comments to author 10.02.24; revised version received 10.03.24; accepted 13.06.24; published 25.07.24.

Please cite as:

Huang J, Guo P, Zhang S, Ji M, An R

Use of Deep Neural Networks to Predict Obesity With Short Audio Recordings: Development and Usability Study

JMIR AI 2024;3:e54885

URL: <https://ai.jmir.org/2024/1/e54885>

doi: [10.2196/54885](https://doi.org/10.2196/54885)

PMID:

©Jingyi Huang, Peiqi Guo, Sheng Zhang, Mengmeng Ji, Ruopeng An. Originally published in JMIR AI (<https://ai.jmir.org>), 25.07.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR AI, is properly cited. The complete bibliographic information, a link to the original publication on <https://www.ai.jmir.org/>, as well as this copyright and license information must be included.

Original Paper

Evaluating Literature Reviews Conducted by Humans Versus ChatGPT: Comparative Study

Mehrnaz Mostafapour¹, PhD; Jacqueline H Fortier¹, MSc; Karen Pacheco¹, MSc; Heather Murray^{1,2}, MD, MSc; Gary Garber^{1,3,4}, MD, FRCPC

¹Canadian Medical Protective Association, Ottawa, ON, Canada

²Department of Emergency Medicine, Queen's University, Kingston, ON, Canada

³Ottawa Hospital Research Institute, Ottawa, ON, Canada

⁴Department of Medicine and the School of Public Health and Epidemiology, University of Ottawa, Ottawa, ON, Canada

Corresponding Author:

Gary Garber, MD, FRCPC

Canadian Medical Protective Association

875 Carling Ave

Ottawa, ON, K1S 5P1

Canada

Phone: 1 800 267 6522

Email: research@cmpa.org

Abstract

Background: With the rapid evolution of artificial intelligence (AI), particularly large language models (LLMs) such as ChatGPT-4 (OpenAI), there is an increasing interest in their potential to assist in scholarly tasks, including conducting literature reviews. However, the efficacy of AI-generated reviews compared with traditional human-led approaches remains underexplored.

Objective: This study aims to compare the quality of literature reviews conducted by the ChatGPT-4 model with those conducted by human researchers, focusing on the relational dynamics between physicians and patients.

Methods: We included 2 literature reviews in the study on the same topic, namely, exploring factors affecting relational dynamics between physicians and patients in medicolegal contexts. One review used GPT-4, last updated in September 2021, and the other was conducted by human researchers. The human review involved a comprehensive literature search using medical subject headings and keywords in Ovid MEDLINE, followed by a thematic analysis of the literature to synthesize information from selected articles. The AI-generated review used a new prompt engineering approach, using iterative and sequential prompts to generate results. Comparative analysis was based on qualitative measures such as accuracy, response time, consistency, breadth and depth of knowledge, contextual understanding, and transparency.

Results: GPT-4 produced an extensive list of relational factors rapidly. The AI model demonstrated an impressive breadth of knowledge but exhibited limitations in in-depth and contextual understanding, occasionally producing irrelevant or incorrect information. In comparison, human researchers provided a more nuanced and contextually relevant review. The comparative analysis assessed the reviews based on criteria including accuracy, response time, consistency, breadth and depth of knowledge, contextual understanding, and transparency. While GPT-4 showed advantages in response time and breadth of knowledge, human-led reviews excelled in accuracy, depth of knowledge, and contextual understanding.

Conclusions: The study suggests that GPT-4, with structured prompt engineering, can be a valuable tool for conducting preliminary literature reviews by providing a broad overview of topics quickly. However, its limitations necessitate careful expert evaluation and refinement, making it an assistant rather than a substitute for human expertise in comprehensive literature reviews. Moreover, this research highlights the potential and limitations of using AI tools like GPT-4 in academic research, particularly in the fields of health services and medical research. It underscores the necessity of combining AI's rapid information retrieval capabilities with human expertise for more accurate and contextually rich scholarly outputs.

(JMIR AI 2024;3:e56537) doi:[10.2196/56537](https://doi.org/10.2196/56537)

KEYWORDS

OpenAIs; chatGPT; AI vs. human; literature search; Chat GPT performance evaluation; large language models; artificial intelligence; AI; algorithm; algorithms; predictive model; predictive models; literature review; literature reviews

Introduction

Artificial intelligence (AI) is a rapidly evolving technology that combines computer programming with large data sets to enable software to perform tasks. Generative AI uses this technology to synthesize content; the system is trained on large volumes of data to identify patterns until it can recognize those patterns and generate novel responses to queries. Large language models (LLMs), such as ChatGPT, are a form of generative AI wherein the software is trained on extensive textual data sets and can generate a response to prompts and questions [1].

AI in general and LLMs in particular are in a period of exponential growth, and researchers are exploring their utility to perform tasks with variable results [1-5]. Previous studies have shown how these tools can help to advance research [4]. One area where there is potential to realize efficiencies is in the creation of literature reviews and syntheses. The pace of scientific publication has been rapidly expanding [6], and AI tools may provide a useful starting point and substantial time savings by automating some elements of a literature search. However, there is little research that compares the results generated using AI with those generated by skilled human researchers.

The purpose of this study is to conduct a literature review using OpenAI's ChatGPT-4 model ("GPT-4") and then conduct a comparative analysis against a review conducted by human researchers.

The way researchers use these tools and optimize the language used to generate a response from ChatGPT, known as prompt engineering, directly impacts the quality of results [7]. Clear, concise, neutral, structured, and specific prompts reduce the model's tendency to respond with generic or off-topic responses, as well as generate an unsubstantiated or false response, also termed an AI hallucination [8]. Therefore, in order to conduct this study, we have introduced an approach to prompt engineering that may assist researchers who wish to use GPT-4 or other LLMs to generate literature reviews.

Textbox 1. The eligibility criteria to identify relevant studies.

Inclusion criteria

- The study described an empirical research study or a literature review.
- The study focused on or described relational factors between physicians and patients impacting patients' satisfaction.
- The record focused on or described the relationship between patients' satisfaction and medicolegal risk against physicians.
- The study focused on or described medicolegal complaints against physicians caused by relational problems between patients and physicians.

Exclusion criteria

- The study was not empirical research, for example, editorials, commentaries, and reports.
- The study contained no explicit mention of physicians and patients' relationships.
- The study was not related to either patients' satisfaction or medicolegal risk against physicians.

Methods

Overview

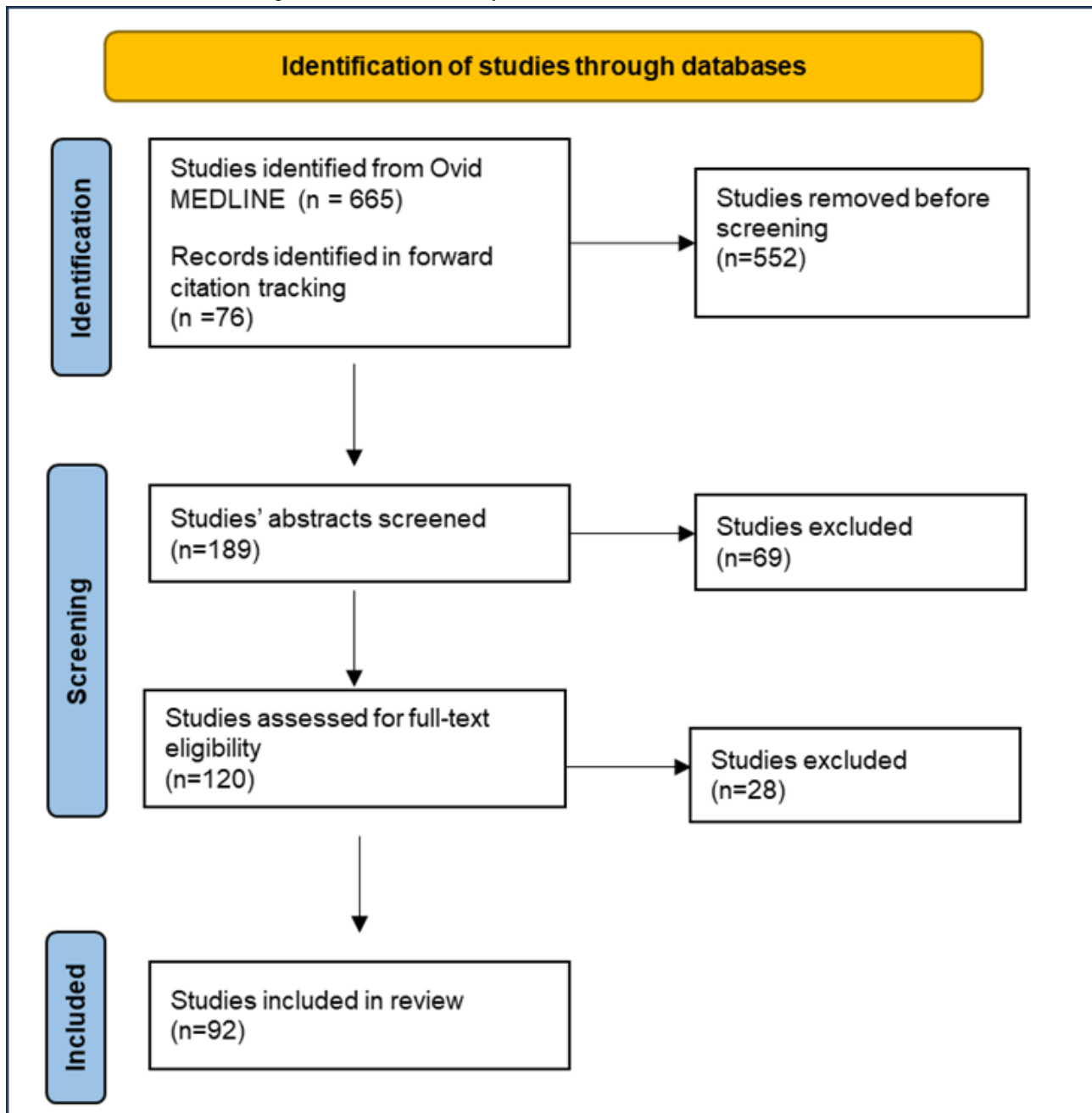
We started with a completed literature review exploring the factors influencing the relational dynamics between the physician and the patient that motivate patients to file medicolegal complaints against physicians [9]. Using this review as a reference standard, we then tasked Open AI's GPT-4 model (training data updated in September 2021) with producing a literature review on the same topic. Subsequently, we compared the results generated by GPT-4 and the literature review conducted by human experts. It should be noted that while GPT-4 was used to generate a literature review and make suggestions for the paper title, it was not used to write this paper.

Human Literature Review

The first author conducted a traditional literature review to identify what factors affect relationships between physicians and patients. They used a systematic approach to ensure transparency and reproducibility. The review included a mix of studies and assessed both qualitative and quantitative data together through thematic analysis [10]. With the help of a research librarian, they developed a search strategy using Medical Subject Headings (MeSH) terms, keywords, and key phrases for a single database (Ovid MEDLINE) to identify articles related to physician-patient relationships. The search strategy was calibrated to identify articles that were most relevant to the research question, rather than prioritizing an approach that would capture every potentially relevant paper (detailed in the "Search strategy developed for literature search led by human researchers" section in [Multimedia Appendix 1](#)).

Subsequently, the librarian screened titles and abstracts, and then the main author screened full-text papers for inclusion against predefined eligibility criteria. Papers had to be empirical research studies or literature reviews that discussed relational factors between physicians and patients that affected patient satisfaction and medicolegal complaints. Studies were excluded if they were not based on empirical research (eg, editorials, commentaries, and reports) or if they were unrelated to the research question ([Textbox 1](#); [Figure 1](#)).

Figure 1. Overview of article screening and inclusion into the study.



We used a thematic analysis approach to review and synthesize the included manuscripts to identify the relational factors that influence patient satisfaction and medicolegal complaints, and we reported the findings in a published narrative review [9].

AI Prompt Engineering for the Literature Search

Previous work suggested that the use of single prompts may not be very effective for complex tasks [2]. We began our process with a single prompt for the literature search (detailed in the “Single Prompt” section in [Multimedia Appendix 1](#)), and the results were clearly inadequate, confirming these findings. Consequently, we developed a series of prompts in an iterative, sequential format. This approach operated on the premise that GPT-4 would benefit from incremental and iterative guidance to yield optimal results. In this approach, the researcher designed sequential prompts based on the assessment of the previous

responses generated by GPT-4, starting from a general prompt, and designing subsequent prompts to refine the output toward the desired form.

An initial series of prompts was used to explore GPT-4’s breadth of knowledge about the factors impacting relationships between physicians and patients. The first prompt was general, simple, and short, asking GPT-4 to list relevant factors to the subject matter. Since we did not know all the relevant factors related to the topic, subsequent prompts were designed to ask for more factors to extend the list of factors and reinforce previous instructions while specifying the desired tone. Further prompts extended the factors and ensured the validity of their content by introducing additional criteria such as the number of sentences and asking for precise references ([Table 1](#); detailed in the “Identifying relational factors” section in [Multimedia Appendix 1](#)).

Table 1. Iterative prompts used to generate contributing factors.

Step	Prompts
1	Write a literature review on the relational problems between physicians and patients that lead to medicolegal complaints against physicians, from a health service researcher perspective, and provide precise references.
2	Please add at least 15 more factors related to relational problems between physicians and patients that lead to medicolegal complaints against physicians to the list, considering the sensitivity, precision, and accuracy of information.
3	You provided me with 21 relational factors between physicians and patients that contribute to the likelihood of filing a medicolegal case against a physician. Please write an elaborated, scientific, and accurate description for each factor that includes at least 15 sentences, and provide at least two real and precise references that support your arguments.

In order to replicate the format of the literature review done by human researchers, the researcher who had conducted the literature review explored a series of prompts to guide GPT-4 through a more in-depth exploration of the identified relational factors. They began by prompting GPT to suggest evidence-based ways to improve each relational factor (eg, “please also describe how to address communication issues using methods derived from scientific publications and research”), which were then evaluated. In cases where the proposed strategy was deemed unsuitable, they either recommended a specific alternative to replace the initial strategy or asked ChatGPT-4 to generate a different one. If the new strategy met the criteria, they instructed GPT-4 to incorporate it into the written description. For example, when asked about

communication issues, GPT-4 first suggested the Four Habit Model [11], which they evaluated to be somewhat out of date. With further prompting, GPT-4 suggested newer strategies to improve communication between physicians and patients, such as the teach-back method and the Shared Decision-Making Model [12], which they then instructed GPT-4 to incorporate in the description. They were able to make these adjustments because they used ChatGPT search while armed with subject matter expertise and an understanding of the available literature for this topic. They leveraged this knowledge to refine the approach to prompt engineering during the process (Table 2; detailed in “Exploring communication as a factor” in [Multimedia Appendix 1](#)).

Table 2. Iterative prompts used to elaborate on each factor.

Step	Prompts
1	Please also describe how to address communication issues using methods derived from scientific publications and research.
2	Is the Four Habit Model the most cited and most recent paper on how to address communication problems? Can you please find a balance between the most cited research papers and the most recent ones, when trying to find references to explain the problem and to address the problem?
3	Please explain the teach-back model and shared decision-making in communication issues using relevant references.

Comparison of Human Versus AI Literature Reviews

To the best of our knowledge, there are no validated tools or checklists to compare human and AI literature reviews. Therefore, we chose to compare the reviews subjectively with respect to the accuracy, response time, comprehensibility, consistency, breadth and depth of knowledge, contextual understanding, and transparency of the outputs. The criteria are defined as follows:

1. Accuracy: we defined the accuracy of the outcome as the percentage of correct responses.
2. Response time: we defined response time as the time it took to conduct the review, including identifying factors and demonstrating what they are.
3. Consistency: we defined consistency as the degree of reliability and stability in the results of a study when it is repeated under similar conditions to ensure it can be replicated.
4. Breadth of knowledge: we defined the breadth of knowledge as the extent and range of information one has access to across various subjects and disciplines.
5. Contextual understanding: contextual understanding refers to the ability to comprehend the meaning and relevance of

information within its specific context. It goes beyond just the literal meaning of words but also the nuances and implications shaped by the situation and the specific circumstances under which the communication occurs [13].

Ethical Considerations

All the studies done at the Canadian Medical Protective Associations have received ethics approval from the ethics review panel of the Advarra Institutional Review Board (Protocol #00020829).

Results

Exploring Literature Search Using AI (GPT-4 Model)

Using iterative prompts and starting from a general prompt, GPT-4 initially generated 6 relational factors. After being repeatedly asked for more factors, it became evident that GPT-4 began to produce unrelated factors after reaching 21, likely due to hallucinations. Overall, GPT-4 generated a list of 21 relational factors and provided 54 references (Table 3). Of these 21 factors, 14% (n=3) were identified as irrelevant. About 24% (n=13) of the references were identified as somewhat related to the topic but not particularly strong in their relevance, and 7.5% (n=4) were identified as irrelevant.

Table 3. Factors affecting the relationships between physicians and patients that lead to patient dissatisfaction and medicolegal complaints as identified in human- and artificial intelligence (AI)-led literature reviews.

Human literature review	AI-generated literature review, following iterative prompts 1-3
Communication	1. Communication issues
<ul style="list-style-type: none"> Understanding patients' concerns and expectations Clarity of communication Information sharing and transparency Tone and attitude 	2. Lack of informed consent
Individual characteristics	3. Perceived negligence or incompetence
<ul style="list-style-type: none"> Physician characteristics Patient characteristics 	4. Mismatched expectations
Perceived care and empathy	5. Perceived lack of care or empathy
Health care system and policies	6. Systemic issues
<ul style="list-style-type: none"> Navigating the health care system Wait times Resource constraints 	7. Cultural and language barriers
	8. Failure to follow-up
	9. Breakdown in continuity of care
	10. Patient autonomy disregarded
	11. Trust erosion
	12. Financial conflicts of interest
	13. Power dynamics
	14. Failure to respect confidentiality
	15. Inadequate documentation ^a
	16. Unaddressed patient concerns
	17. Provider burnout
	18. Poor coordination among care teams ^a
	19. Patient's previous negative experiences
	20. High patient expectations
	21. Medical complexity ^a

^aFactors indicated with an asterisk were identified by GPT-4 but were judged to be inaccurate by human researchers.

GPT-4 demonstrated an impressive ability to retrieve a breadth of information; however, our assessment showed that this information could be superficial, requiring an in-depth investigation to ensure its reliability and validity. Since we were uncertain how many relevant factors were related to the topic, we prompted GPT-4 to extend the list of relevant factors. We also observed that GPT-4 will not communicate to its users when the topic has been saturated or when to stop asking for more information. For example, when we pushed it to go beyond contributing to the relational problems between physicians and patients, GPT-4 provided 30 factors, but the additional factors were increasingly irrelevant or obviously incorrect.

Although the description provided by GPT-4 for each factor was initially short and concise, with prompting, the elaborations for each factor became more detailed and comprehensive. In addition, we noted that GPT-4 initially displayed limitations in adhering to prompted numerical guidelines, such as requesting a specific sentence count, word limit, or number of references, but it started to better follow the instructions when they were reinforced in subsequent prompts.

Our findings showed that GPT-4 can offer relevant responses to questions; however, there were instances where more precise, suitable, or applicable alternative answers existed. For example, when tasked with suggesting a mitigation strategy for communication issues between physicians and patients, GPT-4's initial recommendation was the Four Habits Model. However, upon deeper expert analysis, the researchers determined that the teach-back method and Shared Decision-Making Model

were more fitting for the review. This underscores that initial responses from GPT-4, although relevant, may require further evaluation to determine their optimal relevance and applicability.

Human Literature Review

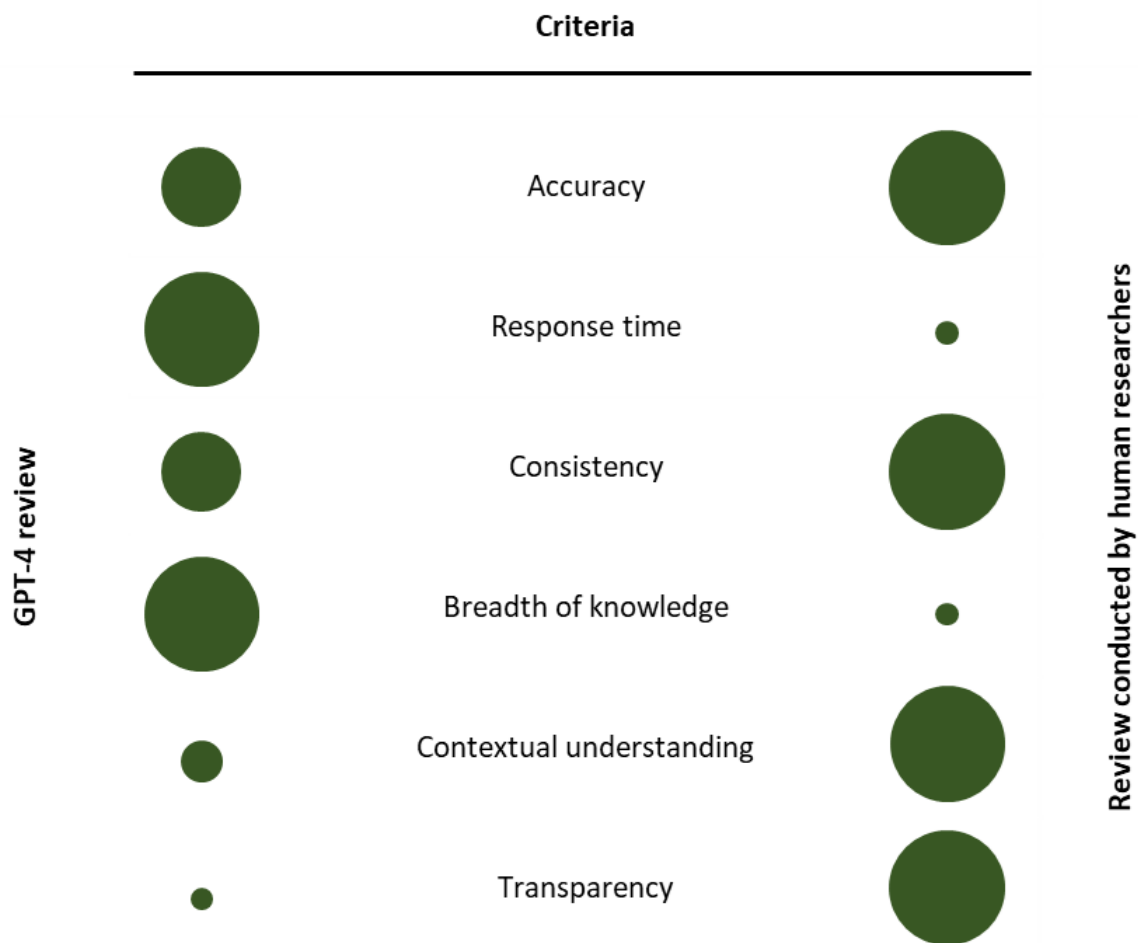
A total of 120 articles were identified for review. Title and abstract screening against the eligibility criteria yielded 113 papers that were directly relevant to our objectives, of which 92 were included for full-text reading and analysis. Two researchers (MM and JHF) reviewed the included articles and identified factors that affected the physician-patient relationship in ways that contributed to patient satisfaction, dissatisfaction, and potential medicolegal complaints. These factors were sorted into the themes and subthemes displayed in [Table 3](#).

Comparison of Human Versus AI Literature Reviews

Overview

While both reviews identified factors influencing the relational problems between physicians and patients, there were important differences. In the human-led literature search, we used a semistructured approach to find relevant references, then conducted a thematic analysis to group the factors into themes and convey the concepts clearly to the target audience. In contrast, GPT-4 used a proprietary search algorithm to explore the web, find relevant articles, and identify relevant factors. Also, it only followed the instructions to list the factors, so there was no synthesis or grouping of the factors. A qualitative comparison of the categories below can be found in [Figure 2](#).

Figure 2. Qualitative comparison of reviews conducted by GPT-4 versus human researchers. Circle sizes (large to small) qualitatively represent differences in criteria between GPT-4 and human researchers; they are not intended for precise measurement.



Accuracy

Of the 21 relational factors that were produced by GPT-4, 86% (n=18) were assessed to be accurate based on our subjective assessment (Table 3; detailed in the “Identifying relational factors” section in Multimedia Appendix 1). As noted above, GPT-4 will continue to suggest factors based on the user’s prompting, so the accuracy rate would decline if the user kept asking for additional factors.

In contrast, experienced human researchers have the nuanced judgment to identify the relevant factors and eliminate the ones that are not relevant to the subject matter. Typically, they can offer a coherent rationale to justify their identification of a factor as either relevant or irrelevant to the subject matter. In the review conducted by researchers, all the identified factors are considered quite relevant to the subject matter, and their relevance is supported by scientific evidence.

Response Time

The AI model generated results within seconds, and the entire series of experiments and prompts were conducted over a few days. The human-led literature review was not conducted as a time trial and occurred as part of a researcher’s regular activities over the course of several months. Had the review been conducted explicitly for this study, it would have required substantially more time for human researchers to search the

literature, read and comprehend the papers, and produce results, compared with GPT-4. Our evaluation indicated that OpenAI’s GPT-4 model demonstrated an unparalleled advantage in response time.

Consistency

In general, the GPT-4 model produced reliable responses to prompts, but similar prompts could sometimes result in variable outputs. We observed that shorter and more precise prompts were more likely to yield consistent results, whereas complexity and length in prompts led to more variability in outputs. When conducting literature reviews, human researchers produce fairly consistent results when they have adequate resources (eg, access to a skilled health research librarian for literature search strategies) and follow established techniques (eg, PRISMA [Preferred Reporting Items for Systematic Review and Meta-Analysis] for systematic reviews) [10].

Breadth and Depth of Knowledge

Our experiments demonstrated a considerable breadth of knowledge within the GPT-4 model, significantly surpassing that of human researchers. This was particularly evident when the model almost instantly generated an extensive list of contributing factors to relational problems between physicians and patients, as well as a comprehensive list of potential mitigation strategies for each factor.

While breadth of knowledge is valuable when conducting literature reviews, synthesizing the information derived from such a review requires deep knowledge and the ability to apply, analyze, and evaluate information related to that topic. This is an area where the GPT-4 model fell short, and a human researcher with experience in a specific subject area may have an advantage.

Contextual Understanding

While LLMs are nonsentient and do not understand meaning in a traditional sense, our experiments revealed that GPT-4 was able to produce outputs that included a satisfactory level of contextual information to allow readers to understand and link key concepts. For example, through iterative prompting, the software was able to produce a list of physician-patient relational issues that included factors as varied as power dynamics, provider burnout, medical complexity, and cultural and language barriers. This level of context was improved by iterative feedback and prompting, providing expanded definitions and additional references [13]. However, given the fact that GPT-4 started to hallucinate when asked to generate more factors, we concluded that it did not have a deep contextual understanding to stop generating meaningless outcomes. On the other hand, human researchers possess an understanding of meaning that consistently results in superior proficiency in interpreting and responding to nuanced contextual elements in this literature search, which would prevent such errors.

Transparency

Another area where human researchers have an advantage is transparency. Human researchers can describe their literature search methods, state and rationalize eligibility criteria, explain the inclusion or exclusion of various articles, describe the approaches used in synthesis, and answer specific questions about their methods. There is significantly less transparency in the way that LLMs process prompts, collect information, and generate outputs at this time. Even when prompted to explain how it completed its literature review, GPT-4 will explain broadly that it drew upon diverse training data but cannot provide a full list of the relevant resources it reviewed, and so the backend review process is almost hidden.

Discussion

Overview

Many researchers are considering how AI tools can support their research. As with any new technology, there is a spectrum of uptake from “early adopters” to “stubborn resistors.” This paper explored how a widely available LLM tool, GPT-4, conducts literature reviews and compares the generated outcomes with a similar review conducted by human researchers.

We found that human-generated literature reviews were more transparent, consistent, and accurate, as long as the literature review was approached systematically and the researcher had sufficient experience and expertise in the subject area. In contrast, GPT-4-generated results were much faster, provided an impressive breadth of content, and were reasonably accurate. We also found that the model was often inconsistent in its

outputs and at times generated irrelevant information, especially if forced to generate a certain number of factors.

One of the fundamental differences between the literature review generated by GPT-4 and humans was in terms of contextual understanding. We attribute this difference to one often-cited limitation of LLMs: their status as so-called “stochastic parrots” [14] that use statistical probabilities of which word is most likely to be next rather than understanding meaning. With prompting, GPT-4 rapidly produced an extensive list of factors that affect the relationship between physicians and patients that appeared very relevant. However, a deeper examination by experts identified inaccurate outputs among accurate ones. This underscores the necessity of expert evaluation in discerning the nuanced veracity of the information generated by GPT-4.

In fact, in this study, we identified 2 potential scenarios where researchers might encounter challenges while working with GPT-4. First, effective communication with the model, specifically through adept prompt engineering, is crucial. Inadequate or improper prompting, particularly for complex tasks like conducting a literature review, leads to unsatisfactory results (detailed in the “Single prompt” section in [Multimedia Appendix 1](#)). Second, novice researchers, unfamiliar with a specific field, might use effective prompting techniques and obtain a broad array of information. This breadth of knowledge can be initially impressive, yet it is important to recognize that the generated content may include errors or inaccurate information. It is for this reason that researchers must carefully review the results to identify and correct potential inaccuracies. The importance of expert oversight in evaluating the reliability of GPT-4-generated content is clear.

This paper introduces an iterative algorithm to effectively search the literature to address the first challenge. We suggested an approach to prompt engineering that uses directive iterative prompts to guide GPT-4 to develop a literature review for researchers. This structured approach includes 2 phases. In the initial phase, researchers are advised to formulate a sequence of prompts that is broad yet precise, progressively becoming more specific. This approach should be designed to incrementally introduce and reinforce instructions, guiding GPT-4 toward generating an output that offers a thorough and comprehensive perspective on a particular subject. In the second phase, the researcher can independently query elements, concepts, or factors identified in the first phase to explore these in greater detail. At all phases of the process, the researcher’s own understanding of the subject will shape the prompts and drastically improve GPT-4’s literature review, suggesting relevant ideas and references while guiding the software away from outdated or incorrect concepts.

We suggest approaching GPT-4 as a research assistant who possesses limited contextual expertise and occasionally synthesizes responses entirely to overcome the second challenge. This requires substantial insight and knowledge from the researcher to diligently guard against the so-called “hallucinations” of the software. Such vigilance is crucial, as GPT-4 can produce convincing yet entirely fabricated content and references [2,15].

For this reason, it seems that GPT-4 might be a more useful tool for experienced researchers looking for wide surveys on a particular topic. The human researcher's knowledge and expertise in a specific area allows them to develop appropriate prompts, iterate with the software to refine the outputs, introduce relevant frameworks and key references, and ultimately guide the process toward the desired output with a clear-eyed understanding of the limitations of what is produced. However, it can also offer different benefits to other audiences, including more novice researchers. Leveraging its extensive knowledge base and inhuman quickness, GPT-4 can help newcomers familiarize themselves with the domain under review. The software acts as an information assistant, offering a wide spectrum of knowledge within a defined domain. In addition, for researchers who have few resources or constrained schedules, it can be used to facilitate the literature review process by offering a robust preliminary draft outline, encompassing key concepts that serve as foundational building blocks. Other studies have explored the potential use of GPT-4 and other LLMs for research tasks such as scholarly writing [2,16], medical writing [15,17,18], and systematic reviews [19]. Still, the rapid improvement in generative AI software has also spurred rapid growth in concerns, such as those related to the ethics of ChatGPT as a coauthor [20] or the potential for it to be used to disseminate misinformation and promote plagiarism [4]. As with any nascent technology, transparency around its use will be essential, and caution is perhaps warranted.

Overall, this study clearly demonstrates the potential utility of GPT-4, an LLM, in supporting the conduct of literature reviews, particularly when an iterative feedback approach to prompt engineering is used. The software successfully reviewed the literature, identified several factors relevant to the subject matter, and was able to respond to prompts requesting additional detail and references. In some instances, and for some researchers, the benefits of using GPT-4 for a literature review (including good breadth of knowledge, reasonable accuracy, and an impressive response time) outweigh the identified shortcomings (including some inconsistency, some inaccuracy, and less depth of knowledge). We suggest that our structured approach to prompt engineering may serve as a model for researchers looking to integrate generative AI into their literature searches. Given the detailed assessment of the generated outcomes with human-led reviews, we recommend approaching these models as an assistant rather than a wise professor; researchers relying on GPT-4 to provide them with a full and nuanced understanding of a complex or rapidly-evolving subject do so at their own peril.

Limitations and Future Research

This study has some limitations. Given the iterative nature of our approach to prompting GPT-4, we did not predefine our prompts or methods, and the researcher leading the prompts (MM) had extensive experience in the subject area; these factors undoubtedly influenced our prompts and thus our outcomes. Our approach to comparing the human- and AI-led literature reviews was subjective, exploratory, and qualitative.

We acknowledge the limitations posed by using a single database and using a human-conducted review as the comparison

standard. However, the opaque nature of ChatGPT's search strategy presents challenges in directly comparing search methodologies. These aspects are critical for interpreting our findings and suggest avenues for future research. In addition, while we have detailed GPT-4's prompt strategies in the [Multimedia Appendix 1](#), the proprietary and evolving nature of its algorithm limits a comprehensive methodological comparison. Future research should examine AI capabilities in detecting emerging trends and gaps, enhancing our understanding of its utility and constraints in academic research.

In our methodology for the human literature search, we used thematic analysis, a subjective process influenced by the researchers' expertise and perspectives. We highlight the inherent subjectivity of thematic analysis as a key limitation. Similarly, our review of ChatGPT's capability to conduct literature reviews acknowledges the qualitative and subjective nature of this evaluation. Our aim was to offer insights and guidance for researchers interested in leveraging AI tools like ChatGPT in their research endeavors.

This study's methodology involved the same researcher in both conducting the human literature review and guiding the AI, as well as participating in the team that evaluated the outcomes. While this was intended to leverage the researcher's subject expertise, it introduces a potential bias, as the researcher was not blinded to the results of the human review during the AI evaluation. This could influence the assessment and interpretation of the AI-generated content. Future studies might consider a more diversified evaluation team to further mitigate bias and enhance the objectivity of the findings.

This study is limited to an in-depth examination of the ChatGPT-4 model, providing a detailed understanding of this specific tool's capabilities and limitations in conducting literature reviews on a particular topic. While this focus allows for a precise evaluation of GPT-4, we acknowledge that this technology is evolving very fast, and it may not reflect the performance of other AI tools that are designed to handle similar tasks. Despite this limitation, our work shows the potential of AI to streamline the initial stages of literature reviews. To build on this foundation, future research should compare the effectiveness of various AI models across a broader range of topics, thereby enhancing our understanding of the general applicability of AI-assisted literature reviews.

Moreover, upcoming studies should focus on enhancing prompt engineering methods to further leverage ChatGPT-4's capabilities in conducting literature reviews. Addressing identified limitations, such as improving the depth and contextual understanding of AI-generated reviews, is crucial. Expanding the training data sets of ChatGPT-4 to include more diverse and recent publications could potentially mitigate issues of relevance and accuracy. In addition, investigating the role of AI in identifying emerging trends and gaps within specific research fields, particularly in health services and medical research, would provide valuable insights into the practical applications and limitations of AI in academic research.

Final Notes and Considerations

In incorporating AI such as ChatGPT into academic research, ethical considerations are crucial. There is the potential for bias in AI outputs, reflecting the biases present in the training data. Ensuring transparency about how AI is used, including prompt selection and response interpretation, is vital for replicability and trust. Responsible use of AI requires acknowledging its limitations and not substituting it for human expertise. As AI technologies become more prevalent in research, it is essential to establish ethical guidelines that promote awareness of bias, transparency, and responsible usage. Integrating ChatGPT-4 and similar LLMs into academic research could dramatically change how we conduct studies, particularly literature reviews. This technology could speed up our ability to study extensive

fields, enabling quicker responses to new information or gaps in knowledge. However, it is crucial to remember that the depth of understanding and critical analysis, which are at the heart of academic work, cannot be fully replicated by AI.

The use of LLMs might also make research more accessible, allowing a wider range of voices to contribute to scholarly conversations. Yet, we must navigate this future carefully, paying close attention to ethical concerns like bias in AI outputs and maintaining transparency in AI's role in research processes. As we move forward, the challenge will be to harness AI's power to enhance our work while ensuring that the essence of research, critical thinking, depth of analysis, and human insight remain at the forefront. The potential is vast, but it is also our responsibility to use these tools wisely.

Conflicts of Interest

None declared.

Multimedia Appendix 1

(A) Single prompt. (B) Identifying relational factors. (C) Exploring communication as a factor. (D) Search strategy developed for literature search led by human researchers.

[[DOCX File, 51 KB - ai_v3i1e56537_app1.docx](#)]

References

1. The AI writing on the wall. *Nat Mach Intell* 2023;5(1):1 [FREE Full text] [doi: [10.1038/s42256-023-00613-9](https://doi.org/10.1038/s42256-023-00613-9)]
2. Lingard L. Writing with ChatGPT: An illustration of its capacity, limitations & implications for academic writers. *Perspect Med Educ* 2023;12(1):261-270 [FREE Full text] [doi: [10.5334/pme.1072](https://doi.org/10.5334/pme.1072)] [Medline: [37397181](https://pubmed.ncbi.nlm.nih.gov/37397181/)]
3. Pavlik J. Collaborating with chatgpt: considering the implications of generative artificial intelligence for journalism and media education. *Journalism & Mass Communication Educator* 2023;78(1):84-93 [FREE Full text] [doi: [10.1177/10776958221149577](https://doi.org/10.1177/10776958221149577)]
4. Van Noorden R, Perkel JM. AI and science: what 1,600 researchers think. *Nature* 2023;621(7980):672-675. [doi: [10.1038/d41586-023-02980-0](https://doi.org/10.1038/d41586-023-02980-0)] [Medline: [37758894](https://pubmed.ncbi.nlm.nih.gov/37758894/)]
5. Perkins M, Roe J. Academic publisher guidelines on AI usage: a ChatGPT supported thematic analysis. *F1000Res* 2023;12:1398 [FREE Full text] [doi: [10.12688/f1000research.142411.1](https://doi.org/10.12688/f1000research.142411.1)]
6. Landhuis E. Scientific literature: Information overload. *Nature* 2016;535(7612):457-458. [doi: [10.1038/nj7612-457a](https://doi.org/10.1038/nj7612-457a)] [Medline: [27453968](https://pubmed.ncbi.nlm.nih.gov/27453968/)]
7. Giray L. Prompt engineering with ChatGPT: a guide for academic writers. *Ann Biomed Eng* 2023;51(12):2629-2633. [doi: [10.1007/s10439-023-03272-4](https://doi.org/10.1007/s10439-023-03272-4)] [Medline: [37284994](https://pubmed.ncbi.nlm.nih.gov/37284994/)]
8. Ji Z, Lee N, Frieske R, Yu T, Su D, Xu Y, et al. Survey of hallucination in natural language generation. *ACM Comput. Surv* 2023;55(12):1-38 [FREE Full text] [doi: [10.1145/3571730](https://doi.org/10.1145/3571730)]
9. Mostafapour M, Fortier J, Garber G. Exploring the dynamics of physician-patient relationships: factors affecting patient satisfaction and complaints. *J Healthc Risk Manag* 2024;43(4):16-25. [doi: [10.1002/jhrm.21567](https://doi.org/10.1002/jhrm.21567)] [Medline: [38706117](https://pubmed.ncbi.nlm.nih.gov/38706117/)]
10. Pluye P, Hong QN, Bush PL, Vedel I. Opening-up the definition of systematic literature review: the plurality of worldviews, methodologies and methods for reviews and syntheses. *J Clin Epidemiol* 2016;73:2-5. [doi: [10.1016/j.jclinepi.2015.08.033](https://doi.org/10.1016/j.jclinepi.2015.08.033)] [Medline: [26898706](https://pubmed.ncbi.nlm.nih.gov/26898706/)]
11. Frankel R, Stein T. Getting the most out of the clinical encounter: the four habits model. *J Med Pract Manage* 2001;16(4):184-191. [Medline: [11317576](https://pubmed.ncbi.nlm.nih.gov/11317576/)]
12. DeWalt D, Broucksou K, Hawk V, Brach C, Hink A, Rudd R, et al. Developing and testing the health literacy universal precautions toolkit. *Nurs Outlook* 2011;59(2):85-94 [FREE Full text] [doi: [10.1016/j.outlook.2010.12.002](https://doi.org/10.1016/j.outlook.2010.12.002)] [Medline: [21402204](https://pubmed.ncbi.nlm.nih.gov/21402204/)]
13. Gumperz J. Contextualization and understanding. In: *Rethinking Context: Language as an Interactive Phenomenon*. Berkeley, CA: University of California; 1988.
14. Bender E, Gebru T, McMillan-Major A, Shmitchell S. On the dangers of stochastic parrots: can language models be too big? : Association for Computing Machinery; 2021 Presented at: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency; Virtual Event; 2021; Canada URL: <https://doi.org/10.1145/3442188.3445922>

15. Alkaissi H, McFarlane S. Artificial hallucinations in ChatGPT: implications in scientific writing. *Cureus* 2023;15(2):e35179. [doi: [10.7759/cureus.35179](https://doi.org/10.7759/cureus.35179)] [Medline: [36811129](https://pubmed.ncbi.nlm.nih.gov/36811129/)]
16. Dergaa I, Chamari K, Zmijewski P, Ben Saad H. From human writing to artificial intelligence generated text: examining the prospects and potential threats of ChatGPT in academic writing. *Biol Sport* 2023;40(2):615-622 [FREE Full text] [doi: [10.5114/biolSport.2023.125623](https://doi.org/10.5114/biolSport.2023.125623)] [Medline: [37077800](https://pubmed.ncbi.nlm.nih.gov/37077800/)]
17. Benichou L, ChatGPT. The role of using ChatGPT AI in writing medical scientific articles. *J Stomatol Oral Maxillofac Surg* 2023;124(5):101456. [doi: [10.1016/j.jormas.2023.101456](https://doi.org/10.1016/j.jormas.2023.101456)] [Medline: [36966950](https://pubmed.ncbi.nlm.nih.gov/36966950/)]
18. Doyal A, Sender D, Nanda M, Serrano R. ChatGPT and artificial intelligence in medical writing: concerns and ethical considerations. *Cureus* 2023;15(8):e43292 [FREE Full text] [doi: [10.7759/cureus.43292](https://doi.org/10.7759/cureus.43292)] [Medline: [37692694](https://pubmed.ncbi.nlm.nih.gov/37692694/)]
19. Qureshi R, Shaughnessy D, Gill KAR, Robinson KA, Li T, Agai E. Are ChatGPT and large language models "the answer" to bringing us closer to systematic review automation? *Syst Rev* 2023;12(1):72 [FREE Full text] [doi: [10.1186/s13643-023-02243-z](https://doi.org/10.1186/s13643-023-02243-z)] [Medline: [37120563](https://pubmed.ncbi.nlm.nih.gov/37120563/)]
20. De Angelis L, Baglivo F, Arzilli G, Privitera G, Ferragina P, Tozzi A, et al. ChatGPT and the rise of large language models: the new AI-driven infodemic threat in public health. *Front Public Health* 2023;11:1166120 [FREE Full text] [doi: [10.3389/fpubh.2023.1166120](https://doi.org/10.3389/fpubh.2023.1166120)] [Medline: [37181697](https://pubmed.ncbi.nlm.nih.gov/37181697/)]

Abbreviations

AI: artificial intelligence

LLM: large language model

MeSH: Medical Subject Headings

PRISMA: Preferred Reporting Items for Systematic Review and Meta-Analysis

Edited by K El Emam, B Malin; submitted 18.01.24; peer-reviewed by A Doyal, M Rizvi; comments to author 17.02.24; revised version received 12.04.24; accepted 31.05.24; published 19.08.24.

Please cite as:

Mostafapour M, Fortier JH, Pacheco K, Murray H, Garber G

Evaluating Literature Reviews Conducted by Humans Versus ChatGPT: Comparative Study

JMIR AI 2024;3:e56537

URL: <https://ai.jmir.org/2024/1/e56537>

doi: [10.2196/56537](https://doi.org/10.2196/56537)

PMID:

©Mehrnaz Mostafapour, Jacqueline H Fortier, Karen Pacheco, Heather Murray, Gary Garber. Originally published in JMIR AI (<https://ai.jmir.org>), 19.08.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR AI, is properly cited. The complete bibliographic information, a link to the original publication on <https://www.ai.jmir.org/>, as well as this copyright and license information must be included.

Original Paper

Leveraging Temporal Trends for Training Contextual Word Embeddings to Address Bias in Biomedical Applications: Development Study

Shunit Agmon¹, MSc; Uriel Singer¹, PhD; Kira Radinsky¹, PhD

Department of Computer Science, Technion—Israel Institute of Technology, Haifa, Israel

Corresponding Author:

Shunit Agmon, MSc

Department of Computer Science

Technion—Israel Institute of Technology

CS Taub Building

Haifa, 3200003

Israel

Phone: 972 73 378 3897

Email: shunit.agmon@gmail.com

Abstract

Background: Women have been underrepresented in clinical trials for many years. Machine-learning models trained on clinical trial abstracts may capture and amplify biases in the data. Specifically, word embeddings are models that enable representing words as vectors and are the building block of most natural language processing systems. If word embeddings are trained on clinical trial abstracts, predictive models that use the embeddings will exhibit gender performance gaps.

Objective: We aim to capture temporal trends in clinical trials through temporal distribution matching on contextual word embeddings (specifically, BERT) and explore its effect on the bias manifested in downstream tasks.

Methods: We present TeDi-BERT, a method to harness the temporal trend of increasing women's inclusion in clinical trials to train contextual word embeddings. We implement temporal distribution matching through an adversarial classifier, trying to distinguish old from new clinical trial abstracts based on their embeddings. The temporal distribution matching acts as a form of domain adaptation from older to more recent clinical trials. We evaluate our model on 2 clinical tasks: prediction of unplanned readmission to the intensive care unit and hospital length of stay prediction. We also conduct an algorithmic analysis of the proposed method.

Results: In readmission prediction, TeDi-BERT achieved area under the receiver operating characteristic curve of 0.64 for female patients versus the baseline of 0.62 ($P < .001$), and 0.66 for male patients versus the baseline of 0.64 ($P < .001$). In the length of stay regression, TeDi-BERT achieved a mean absolute error of 4.56 (95% CI 4.44-4.68) for female patients versus 4.62 (95% CI 4.50-4.74, $P < .001$) and 4.54 (95% CI 4.44-4.65) for male patients versus 4.6 (95% CI 4.50-4.71, $P < .001$).

Conclusions: In both clinical tasks, TeDi-BERT improved performance for female patients, as expected; but it also improved performance for male patients. Our results show that accuracy for one gender does not need to be exchanged for bias reduction, but rather that good science improves clinical results for all. Contextual word embedding models trained to capture temporal trends can help mitigate the effects of bias that changes over time in the training data.

(JMIR AI 2024;3:e49546) doi:[10.2196/49546](https://doi.org/10.2196/49546)

KEYWORDS

natural language processing; NLP; BERT; word embeddings; statistical models; bias; algorithms; gender

Introduction

Background

Word embeddings are machine-learning models that aim to represent words as real numbered vectors. To train the

embeddings, a large text corpus is needed. Contextualized word embeddings such as BERT [1], where the representation of a word depends on its surrounding words, have an immense impact on performance in various natural language processing (NLP) tasks. In the clinical domain, embeddings pretrained on clinical texts can be used to perform biomedical NLP tasks [2]

or predict clinical outcomes for patients [3]. However, if the training corpus contains biases, they may be perpetuated by the embedding model, and affect the performance on downstream tasks [4-6]. Zhang et al [3] show that word embeddings trained on clinical texts cause performance gaps for different genders and races on clinical tasks.

Clinical trials are the main method to evaluate the efficacy of new treatments on patients, but they may contain biases [7]. For decades, clinical trials excluded women participants [8,9]. The reported reasons for this exclusion include uncertainty about the effects of the menstrual cycle on trial results [10] and tragedies that occurred during trials. For instance, after the thalidomide clinical trial, women of childbearing age were excluded from early-phase clinical trials [8]. Underrepresentation of women leads to a misunderstanding of how women respond to various drugs, which ultimately leads to more adverse drug reactions than in men [11-13]. To mitigate such phenomena, in 1993 the US Food and Drug Administration mandated the inclusion of women in trials [8]. Nevertheless, unequal representation of women persists. Clinical results are not well analyzed nor reported for the influence of gender [9,14].

However, women's representation in clinical trials significantly improves over time due to constant social and legislative efforts [8]. In a comprehensive study of over 43,000 clinical trial papers from PubMed [9], the representation of women in 11 disease categories was analyzed. They found that the number of women participants from before 1993 until 2018 grew in 6 categories and was unchanged in 3 more. In the remaining 2 categories, the female participant proportion was traditionally higher than the female prevalence—the proportion of female patients out of all patients with the disease. The decrease indicates that the proportion grew closer to the actual female prevalence. They find that in all the categories combined, women's representation became more accurate. As women's representation improves, discoveries can be less biased toward women, as reflected in changes in relations between concept embeddings over time (Multimedia Appendix 1).

Related Work

Existing methods to remove representational gender bias from word embeddings aim to remove sensitive information, for example, gender, from the embeddings using data augmentation [15,16], in-training methods modifying the training objective [17], or posttraining methods such as projections to subspaces [4,18,19]. Recently, adversarial training [3,20,21] was also applied to remove information about protected attributes, for example, gender or race, from the representations. These methods aim for a notion of fairness named demographic parity [22]: an independence between a model's prediction and the protected attribute. Indeed, a decision model cannot use the protected attribute if it is not recoverable from the embeddings.

However, in the clinical domain, demographic parity should not be applied, since the sensitive attribute (eg, gender) is an important feature in clinical prediction tasks. Therefore, unlike previous works about adversarial debiasing, we do not remove gender information from the embeddings. Instead, we harness the temporal trend of women's inclusion that exists in the corpus

of clinical trials to improve the information captured in the embeddings regarding women.

Another relevant work [23] explored a method where abstracts were weighted by the number of women who participated in the trial to train gender-sensitive Word2vec [24] embeddings. In this work, we aim to explore the benefits of the improvement in female inclusion over time as an alternative method for debiasing. We compare our work to the method in the study by Agmon et al [23] in Multimedia Appendix 2.

The term “temporal distribution matching” was recently used [25] in an entirely different context: time series forecasting, where given a series of samples and their labels over time, a function from samples to labels is learned. Temporal distribution matching in the context of time series forecasting is a method to handle temporal covariate shifts that harm the performance of the learned prediction model. The method is composed of two phases: (1) detecting the different time periods through “temporal distribution characterization” and (2) performing distribution matching on the hidden states of a recurrent neural network model which is the prediction model. To perform the distribution matching, a loss term is added to the model optimization, based on a pairwise distance between the hidden states of the recurrent neural network after consuming each time period of the series. There are 2 main reasons why this method is not applicable to our problem. First, the task is inherently different: we are interested in learning a word representation model, which is an unsupervised task, while the study by Du et al [25] focuses on time series forecasting, which is a supervised task that requires labels. Second, to calculate a loss term such as was introduced in the study by Du et al [25] requires comparing the state of an embedding model after reading all texts from each time period; embedding models usually do not support such a long context in a meaningful way. Instead, our method uses an adversary component to perform the distribution matching while only looking at 1 abstract at a time. Our method can be viewed as an adjustment of temporal distribution matching to the task of word representation learning.

Goal of This Study

One method to use the improvements in clinical trial practices is to repeat past clinical trials using the new practices. However, it is not a feasible option due to both ethical concerns and the costs of clinical trials. From the machine learning point of view, a naive solution would be to train the embedding model only on the more recent papers; but such a model is trained on far less data. This may yield to suboptimal performance on downstream tasks. We aim to train word embeddings that (1) make use of the entire data set of clinical trial abstracts, (2) harness the positive temporal trends in clinical trials, and (3) achieve high performance on the downstream tasks for the underrepresented group.

Intuitively, we would like to match the distribution of earlier clinical findings to that of more recent findings. We present TeDi-BERT—a temporal distribution matching training method, applied to BERT word embeddings. In this method, in parallel with the original training process of the embeddings, an adversarial temporal classifier tries to distinguish old from new samples based on their embeddings, while the embedding model

tries to *decrease* the adversary's performance. Intuitively, if the temporal classifier's performance is low, then the embeddings of older clinical trials are similar to those of more recent clinical trials. The competition between the embedding model and the temporal classifier acts as a temporal distribution matching mechanism. We use the adversarial component because adversarial models were successfully applied in domain adaptation [26], which is similar to our setting: the different time periods can be viewed as 2 domains.

While there are methods to tackle model biases directly, in this work we explore the effects of temporal distribution matching on bias. Additionally, the proposed method can capture a wide range of trends, such as the emergence of new diseases and new practices. However, in this work, we focus on evaluating its effects on gender bias. Although the method is generic, gender bias is a real practical problem, where temporal trends have been present for years [9]. Evaluating other aspects of temporal distribution matching is left for future work.

We evaluated the model on several tasks, including clinical tasks, based on the MIMIC-III data set [27], and compared the performance on female and male patients.

We contributed our code and data sets [28] to the community to be leveraged for additional tasks where subpopulations are underrepresented.

Methods

Overview

A word embedding is a mapping from words to real numbered vectors, such that the vector captures the meaning of the word. Word embeddings are usually trained on a large corpus of text, using a semantic task. For example, in BERT [1] embeddings, some words in the sentence are masked, and the word vectors of the remaining words are used to predict the masked words. The loss from this prediction task is then used to tune the word vectors: the word representations are modified to better perform the task.

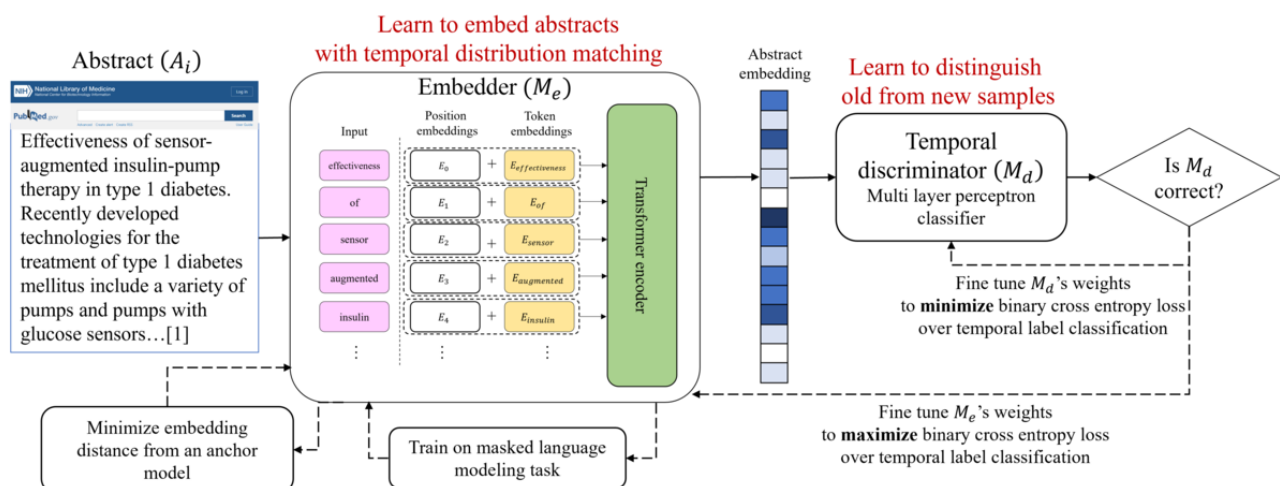
In this work, we describe TeDi-BERT, a temporal distribution matching training method, applied to BERT. We trained the word embeddings on PubMed abstracts of clinical trials between 2010 and 2018. We focused on this time range because there were much fewer clinical trials in ClinicalTrials.gov before 2010, and we used ClinicalTrials.gov to filter the clinical trial abstracts.

One could argue that a better data set to use for training is EHR data, such as the medical notes from MIMIC-III. Numerous factors contributed to our decision not to pursue that course of action. The first is a technical reason: the timestamps available in MIMIC-III were randomly shifted to preserve patient privacy, so visits from different patients are not guaranteed to be in the correct order. Second, the practices and methods in these medical notes represent the conventions used in a single place of medical care, unlike clinical trials which are more diverse, and cover practices and methods from different geographic places. Finally, to validate our choice of training data set, we conducted a qualitative analysis of the trends that exist in clinical trial abstracts and found several examples of real-world trends that were quickly reflected in clinical trial abstract data (Multimedia Appendix 3).

To harness the temporal trends in these clinical trials, we require that the distribution of embeddings of the older abstracts be similar to that of newer abstracts. In addition to training the embedding model on the original semantic task, we simultaneously train it on a temporal classification task.

The abstracts were divided into old, 2010-2013, and new, 2016-2018 (see below for details on the choice of time ranges), and assigned a temporal label. A temporal discriminator, namely, a classifier, aims to distinguish old from new abstracts based on their embeddings. The embedding model, however, aims to *reduce* the classifier's performance by tuning the embeddings. To translate this idea into an architecture (Figure 1), we leveraged the well-received framework of generative adversarial networks (GANs) [29], where 2 components (a generator and a discriminator) compete on a task with opposite goals.

Figure 1. Schematic drawing of the TeDi-BERT model for health care embeddings. Clinical trial abstracts are embedded using a BERT model, and a discriminator aims to distinguish between old and new abstracts. The embedder simultaneously trains on the original embedding task of masked language modeling and regulates the embeddings to resemble an anchor model. TeDi-BERT: temporal distribution matching applied on BERT.



For example, an abstract from 2010 is transformed into a vector representation using the BERT embedder. The embedding vector is fed to the temporal discriminator. Assume that the discriminator correctly predicted that this sample is “old” with probability p . The discriminator’s weights are then updated so that p is closer to 1, while the embedder’s weights are updated so that p is closer to 0.

The embedding model (M_e) is given an abstract, performs the semantic prediction task on the abstract text, and computes the semantic loss (L_{MLM}). Additionally, the same embedding model acts as the generator in the GAN and emits an embedding for the full abstract.

The abstract embedding is fed to the temporal discriminator (M_d), which is a classifier trying to distinguish whether the embedding belongs to a new or old abstract. A binary cross entropy loss (L_{adv}) for this task is computed using the discriminator output and the temporal label. The discriminator aims to minimize this loss. However, the generator aims to both maximize the loss and simultaneously minimize L_{MLM} .

Consider a trivial generator that outputs the same embeddings regardless of the input text. In this case, the discriminator cannot distinguish old from new texts, and L_{adv} would be minimized. To prevent such cases, we wish the model to preserve the original semantics of the texts. We therefore added another term to the loss function, which was meant to anchor the embedding model, so that it did not drift too far from the original embedding. We embed each sample using a frozen anchor model and compute the loss term (L_A) as the L2 Frobenius norm distance between the frozen embedding and the generator’s embedding. The final objective function is given by:



Where θ_M denotes the parameters of a model M , and λ_{adv} and λ_A are hyperparameters used to balance the different components.

Implementation Details

The corpus of clinical trial abstracts from 2010 to 2018 was divided into old (2010-2013) and new (2016-2018) clinical trials. The guiding principle in choosing these time ranges is to create a gap between the 2 time periods, while maintaining a large enough and balanced number of abstracts in each set. The first time range is 1 year longer since there are less abstracts per year in 2010-2013 (~5000 on average) versus 2016-2018 (~9000 on average). The gap is needed for the discriminator task: it is harder to distinguish between abstracts from consecutive years since the temporal trends are slow. When comparing the 2 time ranges, we observed a statistically significant increase over time in the percentage of women participants in clinical trials (Multimedia Appendix 1). This is consistent with previous findings [9] over slightly different time ranges: the total enrollment bias for women was improved from before 1993 (-0.11) to 2014-2018 (-0.05).

As the embedding model, we chose BERT [1], a transformer-based model for contextualized word embeddings.

We used a small version of BERT, named BERT-tiny [30], with 2 transformer layers and a hidden representation size of 128, pretrained on BookCorpus [31] and the English Wikipedia. Smaller models require less computation resources and are therefore more affordable and accessible. Rosin et al [32] have shown that BERT-tiny-based models were comparable to BERT-base in their ability to learn temporal trends. We witnessed a similar phenomenon on the clinical task of length of stay (LOS) prediction (Multimedia Appendix 4).

We initialized the model from a version of BERT which was not trained on any scientific or medical data, so that we could attribute the medical knowledge accumulated in the model only to the clinical trial abstracts in the corpus used in the train set.

As each abstract is long, and BERT has a maximal input length of 512-word pieces, we split it into sentences using the Natural Language Toolkit tokenizer [33]. The generator embeds each sentence. The first m sentence embeddings are concatenated and fed to the discriminator, which is a linear classifier. Hence the classifier size is $d \cdot m + 1$. As 96.97% (21123/21784) of abstracts had up to 20 sentences, we set $m = 20$ and padded shorter abstract embeddings with zeros before feeding them to the discriminator. As a frozen anchor model, we used a BERT model of the same architecture as the generator, initialized similarly but trained only with masked language modeling (MLM) on all of the abstracts.

The embedder and discriminator components of TeDi-BERT were trained simultaneously, 1 batch at a time for 20 epochs. Each component was optimized using the Adam optimizer with a learning rate of $2e-5$. Additional technical details are given in Multimedia Appendix 5.

The TeDi-BERT model used in our experiments was trained with $\lambda_{adv}=0.3$, $\lambda_A=0.3$, hence the weight of the L_{MLM} term was 0.4. We experimented with $\lambda_{adv}, \lambda_A \in \{0, 0.1, \dots, 0.6\}$ and chose the best combination according to the model’s ability to predict the future semantic relatedness of medical concepts (Section S3 in Multimedia Appendix 6).

Experimental Evaluation Setup

The corpus used to train the embedding models is composed of PubMed [34] abstracts describing clinical trials on humans. To select only those abstracts out of the 90,000 available in PubMed version of 2020, we match each abstract with an entry from ClinicalTrials.gov [35] according to the NCT identifier inside the abstract text, leaving 21,784 abstracts, 12,452 of them from 2010-2013 and 2016-2018. We randomly split the data into 70.51% (8780 abstracts) train and 29.49% (3672 abstracts) test, and kept this partition fixed throughout our experiments.

For our downstream tasks, we used 2 different clinical prediction tasks, created based on the MIMIC-III data set [27], an anonymized and publicly available data set that contains information about patients at a massive tertiary care hospital. The data set contained 58,976 hospital admissions with 61,532 intensive care unit (ICU) stays over 46,520 distinct patients. After removing patients aged younger than 18 years (as performed in the study by Lin et al [36]), 38,552 patients remained. We randomly divided the patients into train and test

sets, so that data from a single patient could not appear in both the train and the test. The train set contained 30,817 patients, out of which 43.97% (n=13,553) were female, and the test set contained 7735 patients, out of which 43.33% (n=3352) were female.

Downstream Tasks

LOS prediction—a regression task predicting a patient's LOS in the hospital in days. Predicting LOS is a common clinical task, which is important in hospital resource allocation planning. The predictions can also be taken as indications of the severity and need for different levels of care and recovery.

To predict the LOS we used the patient's diagnoses from their previous admissions, and the primary diagnosis from the current admission, along with demographic features and summary features (number of previous admissions, procedures and diagnoses, and time since the last admission).

Readmission prediction—a classification task predicting unplanned ICU readmission of a patient, at the time of their discharge. Such readmissions indicate an unexpected deterioration in the patient's state. Detecting such cases in advance can improve the quality of care for the patients by allocating special programs and resources that address reasons for readmission. We followed Lin et al [36] for the definition of unplanned readmission: patients that were transferred from the ICU to low-level wards or discharged, but returned to the ICU or died within 30 days. The features used in this prediction task are the patient's diagnoses from previous admissions, and diagnoses and medications from the current admission (which are known at the time of discharge), along with demographic features.

Compared Models

We compared the following models in our experiments:

Nonmedical BERT—a pretrained BERT on English Wikipedia and BookCorpus, not trained on any clinical data [30].

Medical BERT 2010-2018—this baseline represents the natural way to train BERT for clinical uses: training BERT with the MLM task over the clinical texts. The model was initialized with nonmedical BERT and trained for 40 epochs on clinical trial abstracts between 2010 and 2018.

Null it out [18]—As an example of a debiasing method aiming to remove gender information from the embeddings, we applied the method presented in the study by Ravfogel et al [18] on medical BERT 2010-2018. This method was found to be best at debiasing BERT embeddings to remove gender stereotypes [37]. The method is based on iterative null space projection of the embeddings so that the sensitive information (gender) cannot be recovered from them by a linear model. Using the vocabulary of all diseases and drugs used in the clinical tasks data sets, we sampled the 2500 most feminine and 2500 most masculine words, based on their relation to the he-she vector, to build a training and test set for the iterative method. We applied the projection process for 35 iterations. Before the process, a linear classifier could determine the gender of the words in the test set with an accuracy of 0.93. The accuracy dropped to 0.37 after the process.

TeDi-BERT—the TeDi-BERT model, trained as described in the Implementation Details section.

Ethical Considerations

All data sets used in this study are previously existing data sets, which are either anonymous or deidentified. The data sets containing clinical trial information (PubMed and ClinicalTrials.gov) are anonymous: they do not contain any single patient data, only aggregated data from all trial participants. The publicly available MIMIC-III data set that we use is deidentified and was approved as part of the original MIMIC-III project [27] by the institutional review boards of Beth Israel Deaconess Medical Center and the Massachusetts Institute of Technology. Therefore, this research did not require additional approval from an ethics committee.

Another ethical consideration is the use of abstracts in the later time range as reference in the optimization function, although they may still contain biases. This may lead to the model having lower performance on diseases where women are still understudied. However, the results described in the next section show improved performance of our method for women, leading us to believe that while this solution is not flawless, it is a step in the right direction toward addressing the effects of bias in clinical word embeddings. More on this in the Limitations section.

Results

Hospital LOS Regression

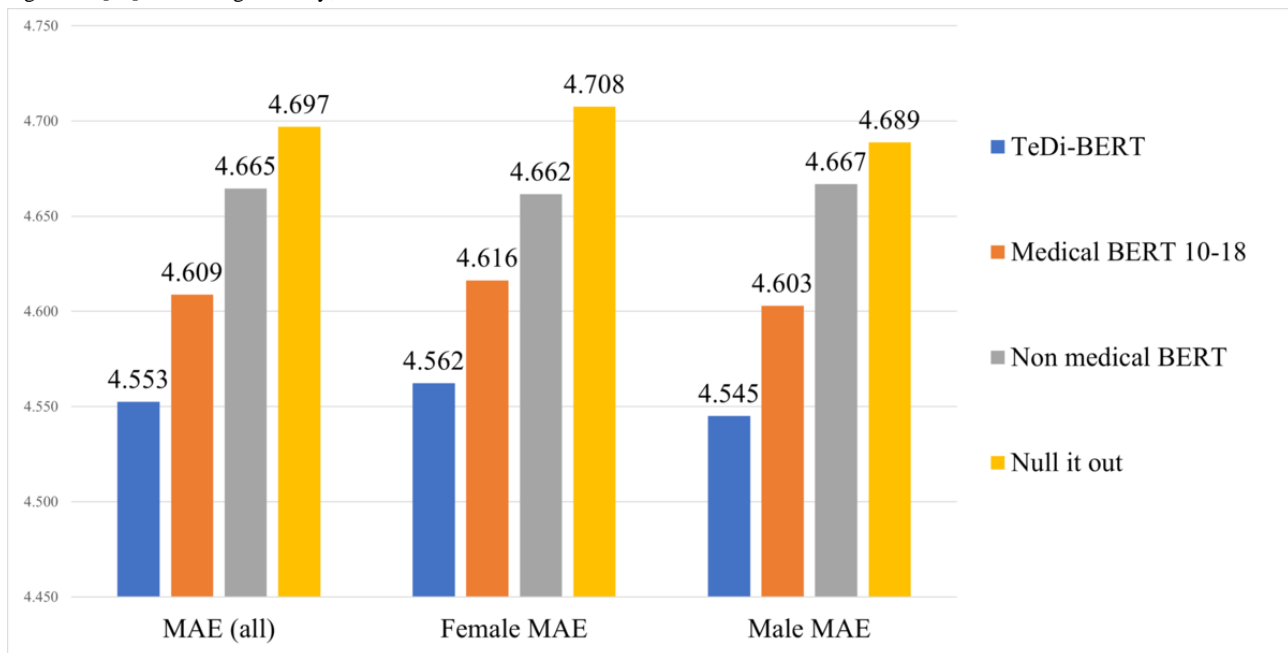
The patient's diagnoses are given as *ICD-9 (International Classification of Diseases, Ninth Revision)* codes and mapped into textual descriptions. The sequence of previous diagnoses is embedded using the evaluated embedding model and aggregated using a long short-term memory network (LSTM) layer. The current diagnosis embedding is concatenated to the LSTM output, and demographic features are added. The combined feature vector is fed into a regression model—a 2-layer neural network. The embedding model is frozen, and only the regression model is allowed to train. As the loss function, we use mean square error in the training process and train each model using the Adam optimizer with a learning rate of $1e-3$ for 10 epochs (after that, the loss increases).

We report the mean absolute error (MAE) for the compared models, calculated over the entire test set, and aggregated by patient gender (Figure 2 [18]). As expected, the nonmedical BERT does not perform well, as it is not tuned on clinical data. Medical BERT trained in the 2010-2018 range reached better results but applying iterative nullspace projection over medical BERT had lower performance than nonmedical BERT. This can be because the projection alters the embedding space, in the effort to remove gender information; these changes may have harmed the semantic information captured in the embeddings. TeDi-BERT performed best, with a significant improvement in MAE for women and for men (Diebold and Mariano [38] test with mean absolute deviation criterion had P value of $<.001$ for both populations). Further analysis by patient ethnicity (Multimedia Appendix 7) shows that TeDi-BERT performed better than medical BERT over all ethnicity groups

but had a specifically large improvement over female patients in minority groups. This suggests that the trends of including underrepresented populations in clinical trials led to the accumulation of a wider knowledge base on these groups. Our model can harness this trend to reach better prediction accuracy

on female patients without harming the accuracy on male patients, and even more so in cases of complex bias types, such as gender and race combined. We hypothesize that the performance improvement for men stems from better conduction of clinical trials with relevance to LOS prediction.

Figure 2. Mean absolute error for LOS regression task using different embeddings. Lower numbers indicate better results. “Null it out” is the work of Ravfogel et al [18]. LOS: length of stay; MAE: mean absolute error.



ICU Readmission Prediction

Each element in each of the medications, diagnoses, and previous diagnoses sequences is embedded using the evaluated embedding model. We aggregate the embeddings using an LSTM (with shared weights over the 3 feature sequences). The concatenation of the aggregated embeddings is fed into a classification model (a 2-layer neural network). The models were trained for 4 epochs using the Adam optimizer with a learning rate of $1e-5$. The results are measured in area under the receiver operating characteristic curve.

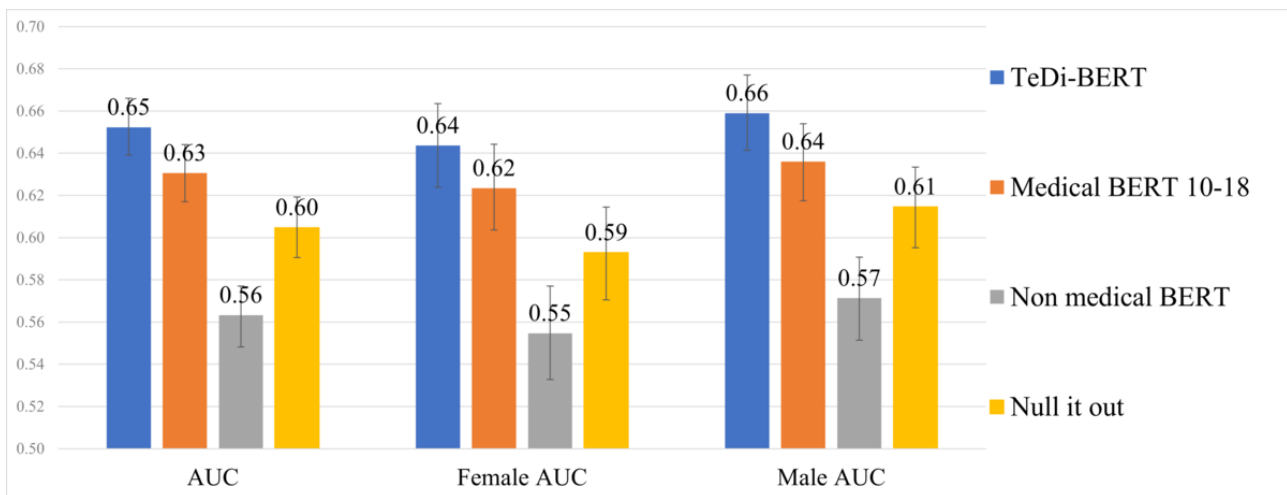
In Lin et al [36], the best model achieved an area under the receiver operating characteristic curve of 0.79, with additional features from the patient medical record events. However, we purposely limited the classifier’s input features to the aforementioned textual fields, since we aim to evaluate the embeddings, and not fully solve the prediction task.

We analyzed the performance of each model per patient gender (Figure 3 [18]). Further, 95% CIs were calculated using bootstrapping with 2000 resamples over the test set. We further validated the significance of the differences using the DeLong test [39]. All differences for all patient groups were significant with $P < .001$.

As in the previous task, nonmedical BERT results were lower than medical BERT and TeDi-BERT. In this task, applying the debiasing method from Ravfogel et al [18] over medical BERT harmed the performance, but it remained better than nonmedical BERT. TeDi-BERT statistically significantly outperformed all models over female and male patients.

Following the results in the 2 clinical tasks, we conclude that debiasing embeddings through the removal of gender information did not improve the performance on downstream tasks. However, we consistently observe that temporal distribution matching improves performance for female patients.

Figure 3. AUC for readmission within 30 days prediction. “Null it out” is the work of Ravfogel et al [18]. AUC: area under the receiver operating characteristic curve.



Algorithm Analysis

To verify that temporal distribution matching does not harm the semantics learned by the embedding model, we evaluated its quality as a language model. We measured the MLM loss on the validation set of the PubMed corpus (Section S1 in [Multimedia Appendix 6](#)). TeDi-BERT's loss (2.650) was close to that of medical BERT (3.292), indicating that our algorithm maintains the semantic performance of BERT, despite the additional objective of temporal distribution matching. Additionally, we tested the models on named entity recognition tasks (Section S2 in [Multimedia Appendix 6](#)) and found that TeDi-BERT did not harm the performance in this task compared to the medical BERT model.

Next, we compared the models on their ability to predict future semantic relatedness of medical concepts, by ranking pairs of medical concepts according to their embedding similarity in each model and comparing the ranking correlation to that of a medical BERT model trained on 2020 abstracts (Section S3 in [Multimedia Appendix 6](#)). TeDi-BERT reached the highest-ranking correlation, meaning that TeDi-BERT was able to predict concept similarity from 2020 better than medical BERT, without ever training on texts from 2020. This strengthens our hypothesis that indeed TeDi-BERT can better capture temporal trends in the embeddings, as measured by word similarities, compared to other BERT models.

Additionally, we performed an ablation test, to evaluate the impact of the anchor model in TeDi-BERT (Section S4 in [Multimedia Appendix 6](#)). A TeDi-BERT model without an anchor model performed similarly to TeDi-BERT on the MLM task, but its performance on the semantic relatedness task was the lowest of all compared baselines. This shows the necessity of using an anchor model in the training process of distribution matching.

Finally, we used another ablation test to assess the impact of the weight given to old and new abstracts in the training process (Section S5 in [Multimedia Appendix 6](#)). We found that a higher weight given to old abstracts caused lower performance in both clinical tasks and the semantic relatedness task. We concluded

that indeed matching the older abstracts to the new ones has a positive impact on performance.

Comparison to Imbalanced Learning Methods

In the MIMIC-III downstream tasks, one could argue that the unbalanced numbers of female (43.97%, 13,553/30,817) and male patients cause a performance gap. We experimented with 3 methods of handling imbalanced data. In all methods, the training set for both tasks was modified to contain 50% women, without modifying the test set.

- Downsampling—downsampling the male patients randomly so that female and male patient numbers are equal (13,553) in the training set.
- Synthetic Minority Over-Sampling Technique (SMOTE) [40]—a classic imbalanced learning method to generate synthetic samples based on neighbors from the same group. We applied SMOTE on the female patients in each downstream task separately and generated 3711 additional samples, so the train set contained 17,264 male patients and 17,264 female patients.
- MedGAN [41]—a widely used synthetic generation method for patient data, that has recently shown promising results in predictive diagnostic tasks. MedGAN combines an autoencoder and a GAN to generate realistic synthetic patient data. For each downstream task, we trained MedGAN on the female patient admissions in the training set and used it to generate additional synthetic admissions, so the train set contained 17,264 male patients and 17,264 female patients.

We trained our prediction models with medical BERT 2010-2018 embeddings on the modified training sets, using the same methods and parameters as in our main results, and compared the results to TeDi-BERT.

In ICU readmission prediction ([Figure 4](#)), downsampling the male patients harmed the performance for both male and female patients and for both models. SMOTE and MedGAN upsampling improved the performance for both populations and both models, but TeDi-BERT still outperformed medical BERT 2010-2018 under MedGAN ($P=.03$ for female patients, $P=.002$ for male patients) and SMOTE ($P<.001$).

In LOS prediction (Figure 5), downsampling and SMOTE performance, for both patient populations. upsampling harmed medical BERT's and TeDi-BERT's

Figure 4. Readmission prediction—comparison of TeDi-BERT versus medical BERT under various methods of handling imbalanced data. The performance is measured in area under the ROC curve, so higher numbers indicate better results. Further, 95% CIs were calculated using bootstrapping with 2000 resamples over the test set. AUC: area under the receiver operating characteristic curve; ROC: receiver operating characteristic curve; TeDi-BERT: temporal distribution matching applied on BERT.

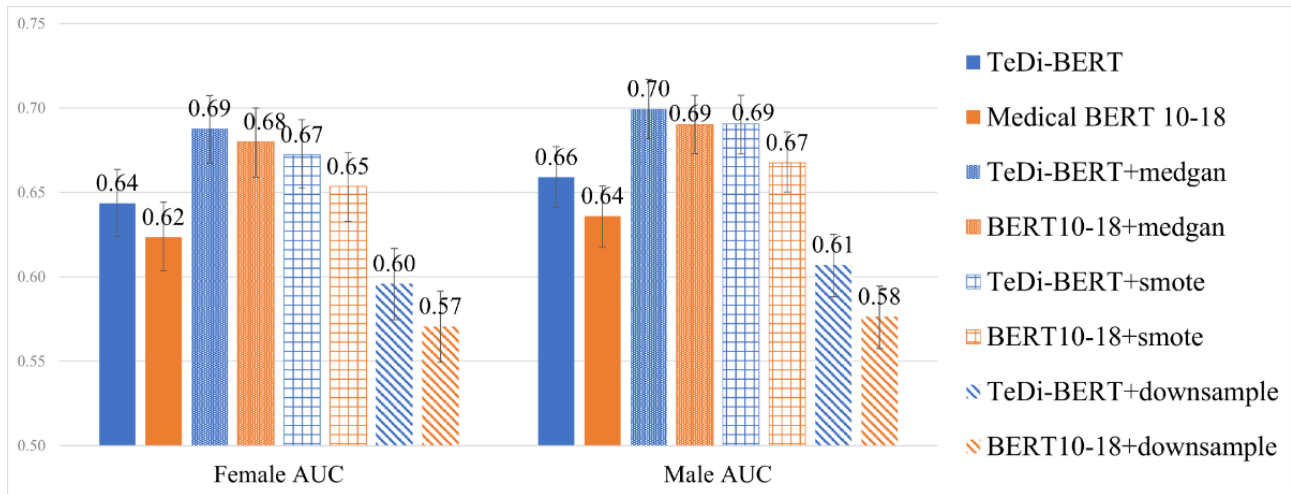
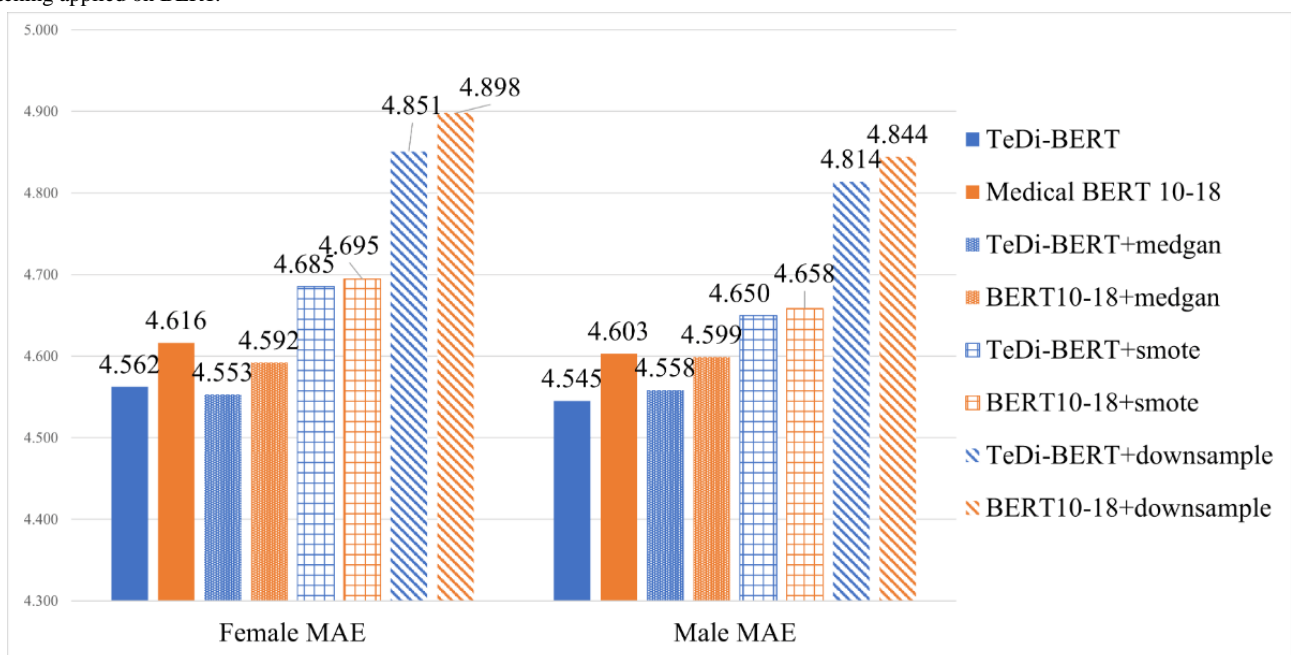


Figure 5. Length of stay prediction—comparison of TeDi-BERT versus medical BERT under various methods of handling imbalanced data. The performance is measured in mean absolute error, so lower numbers indicate better results. MAE: mean absolute error; TeDi-BERT: temporal distribution matching applied on BERT.



MedGAN sampling did not harm the performance, but it did not significantly improve it for either of the models. It is possible that the generated female samples were too noisy to provide added value. Additionally, these methods were designed for much more extreme imbalances than in this setting. This is consistent with several previous works: in multilingual translation [42], upsampling low-resource languages did not robustly improve the loss. In a classification of diseases from textual descriptions of symptoms [43], upsampling rare diseases led to unstable results and in some cases hurt performance.

Over both tested tasks, both populations, and all 3 imbalanced learning methods, TeDi-BERT performed better than medical

BERT 2010-2018. We conclude that imbalanced learning techniques may improve performance, but it is not robust to all tasks and models. As with many other possible techniques to improve performance (data cleaning, feature engineering, etc), imbalanced learning techniques may be applied independently from the choice of embedding model.

Discussion

Principal Results

In both clinical tasks, TeDi-BERT's performance for female patients was significantly improved compared to medical BERT 2010-2018, while improving performance on male patients as

well. This is even though both models were trained on the same data set of clinical trial abstracts. The advantages of the TeDi-BERT method were especially large for population groups subject to intersectional biases ([Multimedia Appendix 7](#)), which suggests that other than gender inclusion, additional improvement trends in clinical trials were captured by the TeDi-BERT model. When analyzing the contribution of our method for different feature types in the LOS task ([Multimedia Appendix 8](#)), we found that for both models, the primary diagnosis was more predictive of the LOS than the previous diagnoses, but TeDi-BERT was able to use the information in previous diagnoses to reduce the MAE more than medical BERT 2010-2018.

A baseline debiasing method based on the removal of gender information from word embeddings [18] did not perform well in the clinical prediction tasks, achieving worse results than medical BERT 2010-2018. This validates our hypothesis that the removal of information about a sensitive attribute from the embeddings is not a suitable strategy for debiasing medical embeddings since that sensitive attribute contains valuable clinical information.

In the semantic task of MLM (Section S1 in [Multimedia Appendix 6](#)), TeDi-BERT's performance surpassed that of medical BERT 2010-2018, despite the competing objective functions of the generator and the discriminator. In another semantic task based on temporal trends (Section S3 in [Multimedia Appendix 6](#)), while both models were trained on the same data set, TeDi-BERT's output was more similar to that of a model trained only on clinical trials from 2020. This validates our hypothesis that TeDi-BERT is better at capturing the temporal trends in the data than medical BERT 2010-2018.

When comparing TeDi-BERT to various imbalanced learning methods, we found that temporal distribution matching had a consistent contribution to performance, while imbalanced learning methods harmed performance in some cases.

When comparing TeDi-BERT to gender-sensitive weighting of the corpus ([Multimedia Appendix 2](#)), we found that gender-sensitive weighting was not a good fit for debiasing BERT embeddings for health care, despite its success for Word2vec embeddings. We hypothesize that this is due to the complexity of the BERT embedding model versus Word2vec and that a finer method is required for debiasing BERT embeddings.

The empirical results show the merit of debiasing embeddings for improving the performance of clinical tasks. Despite the remaining biases in the newer clinical trials, leveraging the temporal trends of bias reduction was successful for the reduction of biases in the embeddings.

Although many works show the trade-off between fairness and accuracy [44-46], our results show that accuracy for one gender does not need to be exchanged for bias reduction, but rather that good science improves clinical results for all.

Limitations

Our work has several limitations. In our TeDi-BERT implementation, we divided clinical trials into 2 time ranges (old and new). This approach is inspired by related work in adversarial domain adaptation [26], where there is a source and target domain. For future work, we wish to expand the approach to a continuous prediction. Additionally, the temporal distribution matching might obfuscate temporal markers such as new diseases or treatments; this can be mitigated by the development of techniques to handle out-of-vocabulary words. Finally, another limitation is the remaining biases in recent clinical trials and the continued underrepresentation of women in them. The use of a still-biased data distribution as the optimization target may cause difficulties in the categorization of diseases where women are still not studied enough, because the knowledge captured in the word embeddings about these conditions may still be partial. However, in many diseases (eg, cardiovascular diseases, anemia, osteoporosis, and more) the situation has greatly improved in recent years. As a result, TeDi-BERT achieved higher performance and lower gender performance gaps in the tested clinical tasks. While it is not a perfect solution, the experimental results show that it is in the correct direction toward fixing the problem. We believe that temporal distribution matching is a good proxy for bias mitigation, but more direct approaches should also be tested.

Conclusions

The use of clinical trials as a training corpus for embedding models should be conducted with care while taking precautions against the long-existing biases in clinical trials. We presented TeDi-BERT, a method for training word embeddings while harnessing a temporal trend in the corpus. The method includes a novel use of the GAN framework to regularize for temporal distribution matching on embedded samples. We implemented our method on BERT, a contextual embedding model that achieved state-of-the-art results in many NLP tasks, and trained it on clinical trial abstracts, where biases, and especially enrollment gender bias, are reduced over time for a significant portion of researched concepts. In our experimental evaluation, we demonstrated performance improvement over BERT in clinical prediction tasks. We found that the performance particularly improved for female patients for all tasks, and for male patients either improved or did not harm performance. This suggests that adjusting for bias in research can benefit clinical results for all patients.

Acknowledgments

SA, USinger, and KR were involved in drafting or critically revising the presented work. SA and USinger designed the model architecture, and KR supervised the methods and their correctness. SA and KR designed the experiments. All authors gave final approval of the submitted paper. Generative artificial intelligence was not used in any part of this paper's writing. This research was partially funded by the student research prize for cross-principal investigator collaboration in data science in funding of the Israeli Planning and Budgeting Committee (VATAT).

Data Availability

The PubMed data is publicly available [34]. The clinical trials metadata is publicly available [35]. MIMIC-III is publicly available [47] pending the completion of a training course.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Analysis of chosen time periods.

[DOCX File , 24 KB - ai_v3i1e49546_app1.docx]

Multimedia Appendix 2

Comparison to gender-sensitive debiasing.

[DOCX File , 24 KB - ai_v3i1e49546_app2.docx]

Multimedia Appendix 3

Qualitative analysis of trends in clinical trials.

[DOCX File , 78 KB - ai_v3i1e49546_app3.docx]

Multimedia Appendix 4

Comparison to larger BERT models.

[DOCX File , 23 KB - ai_v3i1e49546_app4.docx]

Multimedia Appendix 5

TeDi-BERT technical details. TeDi-BERT: temporal distribution matching applied on BERT.

[DOCX File , 21 KB - ai_v3i1e49546_app5.docx]

Multimedia Appendix 6

Algorithm analysis.

[DOCX File , 58 KB - ai_v3i1e49546_app6.docx]

Multimedia Appendix 7

Analysis of length of stay performance by gender and ethnicity.

[DOCX File , 337 KB - ai_v3i1e49546_app7.docx]

Multimedia Appendix 8

Ablation—feature set contribution.

[DOCX File , 23 KB - ai_v3i1e49546_app8.docx]

References

1. Devlin J, Chang MW, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. 2019 Presented at: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers); 2019 June 2-June 7; Minneapolis, Minnesota p. 4171-4186 URL: <https://aclanthology.org/N19-1423.pdf>
2. Lee J, Yoon W, Kim S, Kim D, Kim S, So C, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* 2020;36(4):1234-1240 [FREE Full text] [doi: [10.1093/bioinformatics/btz682](https://doi.org/10.1093/bioinformatics/btz682)] [Medline: [31501885](https://pubmed.ncbi.nlm.nih.gov/31501885/)]
3. Zhang H, Lu A, Abdalla M, McDermott M, Ghassemi M. Hurtful words: quantifying biases in clinical contextual word embeddings. 2020 Presented at: Proceedings of the ACM Conference on Health, Inference, and Learning; 2020 April 2-4; Toronto, Ontario, Canada. [doi: [10.1145/3368555.3384448](https://doi.org/10.1145/3368555.3384448)]
4. Bolukbasi T, Chang KW, Zou JW, Saligrama V, Kalai A. Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. *Adv Neural Inf Process Syst* 2016:9781510838819.
5. Basta C, Costa-jussà MR, Casas N. Evaluating the underlying gender bias in contextualized word embeddings. 2019 Presented at: Proceedings of the First Workshop on Gender Bias in Natural Language Processing; 2019 August 2; Florence, Italy p. 33-39. [doi: [10.18653/v1/w19-3805](https://doi.org/10.18653/v1/w19-3805)]

6. Kurita K, Vyas N, Pareek A, Black AW, Tsvetkov Y. Measuring bias in contextualized word representations. 2019 Presented at: Proceedings of the First Workshop on Gender Bias in Natural Language Processing; 2019 August 2; Florence, Italy. [doi: [10.18653/v1/w19-3823](https://doi.org/10.18653/v1/w19-3823)]
7. Higgins JPT, Altman DG, Gøtzsche PC, Jüni P, Moher D, Oxman AD, Cochrane Bias Methods Group, Cochrane Statistical Methods Group. The Cochrane collaboration's tool for assessing risk of bias in randomised trials. *BMJ* 2011;343:d5928 [FREE Full text] [doi: [10.1136/bmj.d5928](https://doi.org/10.1136/bmj.d5928)] [Medline: [22008217](https://pubmed.ncbi.nlm.nih.gov/22008217/)]
8. Liu KA, Mager NAD. Women's involvement in clinical trials: historical perspective and future implications. *Pharm Pract (Granada)* 2016;14(1):1-9 [FREE Full text] [doi: [10.18549/PharmPract.2016.01.708](https://doi.org/10.18549/PharmPract.2016.01.708)] [Medline: [27011778](https://pubmed.ncbi.nlm.nih.gov/27011778/)]
9. Feldman S, Ammar W, Lo K, Trepman E, van Zuylen M, Etzioni O. Quantifying sex bias in clinical studies at scale with automated data extraction. *JAMA Netw Open* 2019;2(7):e196700 [FREE Full text] [doi: [10.1001/jamanetworkopen.2019.6700](https://doi.org/10.1001/jamanetworkopen.2019.6700)] [Medline: [31268541](https://pubmed.ncbi.nlm.nih.gov/31268541/)]
10. McGregor AJ. Sex bias in drug research: a call for change. *Pharm J, PJ* 2016;296(7887):296(2887) [FREE Full text]
11. Tran C, Knowles SR, Liu BA, Shear NH. Gender differences in adverse drug reactions. *J Clin Pharmacol* 1998;38(11):1003-1009. [doi: [10.1177/009127009803801103](https://doi.org/10.1177/009127009803801103)] [Medline: [9824780](https://pubmed.ncbi.nlm.nih.gov/9824780/)]
12. Zopf Y, Rabe C, Neubert A, Gassmann KG, Rascher W, Hahn EG, et al. Women encounter ADRs more often than do men. *Eur J Clin Pharmacol* 2008;64(10):999-1004. [doi: [10.1007/s00228-008-0494-6](https://doi.org/10.1007/s00228-008-0494-6)] [Medline: [18604529](https://pubmed.ncbi.nlm.nih.gov/18604529/)]
13. Whitley HP, Lindsey W. Sex-based differences in drug activity. *Am Fam Physician* 2009;80(11):1254-1258 [FREE Full text] [Medline: [19961138](https://pubmed.ncbi.nlm.nih.gov/19961138/)]
14. Geller SE, Koch AR, Roesch P, Filut A, Hallgren E, Carnes M. The more things change, the more they stay the same: a study to evaluate compliance with inclusion and assessment of women and minorities in randomized controlled trials. *Acad Med* 2018;93(4):630-635 [FREE Full text] [doi: [10.1097/ACM.0000000000002027](https://doi.org/10.1097/ACM.0000000000002027)] [Medline: [29053489](https://pubmed.ncbi.nlm.nih.gov/29053489/)]
15. Maudslay R, Gonen H, Cotterell R, Simone T. It's all in the name: mitigating gender bias with name-based counterfactual data substitution. 2019 Presented at: Proceedings of the Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP); 2019; Hong Kong, China. [doi: [10.18653/v1/d19-1530](https://doi.org/10.18653/v1/d19-1530)]
16. Lu K, Mardziel P, Wu F, Amancharla P. Gender bias in neural natural language processing. In: *Logic, Language, and Security*. Cham: Springer; 2020:189-202.
17. Zhao J, Zhou Y, Li Z, Wang W, Chang KW. Learning gender-neutral word embeddings. 2018 Presented at: Proceedings of the Conference on Empirical Methods in Natural Language Processing; 2018 October 31 - November 4; Brussels, Belgium p. 4847-4853. [doi: [10.18653/v1/d18-1521](https://doi.org/10.18653/v1/d18-1521)]
18. Ravfogel S, Elazar Y, Gonen H, Twiton M, Goldberg Y. Null It Out: guarding protected attributes by iterative nullspace projection. 2020 Presented at: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020; 2020 July 5-10; Online p. 7237-7256. [doi: [10.18653/v1/2020.acl-main.647](https://doi.org/10.18653/v1/2020.acl-main.647)]
19. Dev S, Li T, Phillips JM, Srikumar V. On measuring and mitigating biased inferences of word embeddings. *Proc the AAAI Conf Artif Intell* 2020;34(05):7659-7666. [doi: [10.1609/aaai.v34i05.6267](https://doi.org/10.1609/aaai.v34i05.6267)]
20. Elazar Y, Goldberg Y. Adversarial removal of demographic attributes from text data. 2018 Presented at: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing; 2018 October 31 - November 4; Brussels, Belgium p. 11-21. [doi: [10.18653/v1/d18-1002](https://doi.org/10.18653/v1/d18-1002)]
21. Zhang BH, Lemoine BM, Mitchell M. Mitigating unwanted biases with adversarial learning. 2018 Presented at: Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society; 2018 February 2-3; New Orleans, LA, USA. [doi: [10.1145/3278721.3278779](https://doi.org/10.1145/3278721.3278779)]
22. Mehrabi N, Morstatter F, Saxena N, Lerman K, Galstyan A. A survey on bias and fairness in machine learning. *ACM Comput Surv* 2021;54(6):1-35. [doi: [10.1145/3457607](https://doi.org/10.1145/3457607)]
23. Agmon S, Gillis P, Horvitz E, Radinsky K. Gender-sensitive word embeddings for healthcare. *J Am Med Inform Assoc* 2022;29(3):415-423 [FREE Full text] [doi: [10.1093/jamia/ocab279](https://doi.org/10.1093/jamia/ocab279)] [Medline: [34918101](https://pubmed.ncbi.nlm.nih.gov/34918101/)]
24. Mikolov T, Chen K, Corrado G, Dean J. Efficient estimation of word representations in vector space. *arXiv Preprint posted online on January 16, 2013*. [doi: [10.48550/arXiv.1301.3781](https://doi.org/10.48550/arXiv.1301.3781)]
25. Du Y, Wang J, Feng W, Pan S, Qin T, Xu R, et al. Adarnn: adaptive learning & forecasting of time series. 2021 Presented at: Proceedings of the 30th ACM international conference on information & knowledge management; 2021 November 1-5; Virtual Event Queensland, Australia p. 402-411. [doi: [10.1145/3459637.3482315](https://doi.org/10.1145/3459637.3482315)]
26. Ganin Y, Ustinova E, Ajakan H, Germain, P, Larochelle H, Laviolette F, et al. Domain-adversarial training of neural networks. *J Mach Learn Res* 2016;17:2096-2030. [doi: [10.1007/978-3-319-58347-1_10](https://doi.org/10.1007/978-3-319-58347-1_10)]
27. Johnson AEW, Pollard TJ, Shen L, Lehman LH, Feng M, Ghassemi M, et al. MIMIC-III, a freely accessible critical care database. *Sci Data* 2016;3:1-9 [FREE Full text] [doi: [10.1038/sdata.2016.35](https://doi.org/10.1038/sdata.2016.35)] [Medline: [27219127](https://pubmed.ncbi.nlm.nih.gov/27219127/)]
28. TeDi-BERT Repository—Temporal Distribution Matching Applied on BERT. URL: <https://github.com/shunita/tedibert> [accessed 2023-05-01]
29. Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, et al. Generative adversarial networks. *Commun ACM* 2020;63(11):139-144. [doi: [10.1145/3422622](https://doi.org/10.1145/3422622)]

30. Google BERT uncased L2 H128 A2. URL: https://huggingface.co/google/bert_uncased_L-2_H-128_A-2 [accessed 2023-02-14]
31. Zhu Y, Kiros R, Zemel R, Urtasun R, Torralba A, Fidler S. Aligning books and movies: towards story-like visual explanations by watching movies and reading book. Santiago, Chile: Proceedings of the IEEE International Conference on Computer Vision (ICCV); 2015 Presented at: Proceedings of the IEEE International Conference on Computer Vision (ICCV); 2015 December 11-18; Santiago, Chile. [doi: [10.1109/iccv.2015.11](https://doi.org/10.1109/iccv.2015.11)]
32. Rosin GD, Guy I, Radinsky K. Time masking for temporal language models. 2022 Presented at: Proceedings of the 15th ACM International Conference on Web Search and Data Mining; 2022 February 21-25; Virtual Event AZ, USA. [doi: [10.1145/3488560.3498529](https://doi.org/10.1145/3488560.3498529)]
33. Bird S, Klein E, Loper E. Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit. US: O'Reilly Media, Inc; 2009.
34. PubMed. URL: <https://pubmed.ncbi.nlm.nih.gov> [accessed 2022-11-01]
35. ClinicalTrials.gov - information on clinical trials and human research studies. URL: <https://clinicaltrials.gov/> [accessed 2022-11-01]
36. Lin Y, Zhou Y, Faghri F, Shaw MJ, Campbell RH. Analysis and prediction of unplanned intensive care unit readmission using recurrent neural networks with long short-term memory. PLoS One 2019;14(7):e0218942 [FREE Full text] [doi: [10.1371/journal.pone.0218942](https://doi.org/10.1371/journal.pone.0218942)] [Medline: [31283759](https://pubmed.ncbi.nlm.nih.gov/31283759/)]
37. Meade N, Poole-Dayane E, Reddy S. An empirical survey of the effectiveness of debiasing techniques for pre-trained language models. 2022 Presented at: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers); 2022 May 22-27; Dublin, Ireland p. 1878-1898. [doi: [10.18653/v1/2022.acl-long.132](https://doi.org/10.18653/v1/2022.acl-long.132)]
38. Diebold FX, Mariano RS. Comparing predictive accuracy. J Bus Econ Stat 2002;20(1):134-144. [doi: [10.1198/073500102753410444](https://doi.org/10.1198/073500102753410444)]
39. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics 1988;44(3):837-845. [doi: [10.2307/2531595](https://doi.org/10.2307/2531595)]
40. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. JAIR 2002;16:321-357. [doi: [10.1613/jair.953](https://doi.org/10.1613/jair.953)]
41. Choi E, Biswal S, Malin B, Duke J, Stewart WF, Sun J. Generating multi-label discrete patient records using generative adversarial networks. 2017 Presented at: Proceedings of the 2nd Machine Learning for Healthcare Conference; 2017 August 18-19; Boston, MA, United States p. 286-305.
42. Li X, Gong H. Robust optimization for multilingual translation with imbalanced data. Adv Neural Inf Process Syst 2021;34:25086-25099.
43. Li X, Wang Y, Wang D, Yuan W, Peng D, Mei Q. Improving rare disease classification using imperfect knowledge graph. BMC Med Inform Decis Mak 2019;19(Suppl 5):238 [FREE Full text] [doi: [10.1186/s12911-019-0938-1](https://doi.org/10.1186/s12911-019-0938-1)] [Medline: [31801534](https://pubmed.ncbi.nlm.nih.gov/31801534/)]
44. Dwork C, Immorlica N, Kalai A. Decoupled classifiers for group-fair and efficient machine learning. PMLR 2018;81:119-133.
45. Kusner M, Loftus J, Russell C, Silva R. Counterfactual fairness. 2017 Presented at: Advances in Neural Information Processing Systems; 2017 December 4-9; Long Beach, CA, United States p. 4069-4079.
46. Menon AK, Williamson RC. The cost of fairness in binary classification. 2018 Presented at: Conference on Fairness, Accountability and Transparency; 2018 February 23-24; New York, NY, United States.
47. MIMIC-III clinical database. URL: <https://physionet.org/content/mimiciii/1.4/> [accessed 2024-08-29]

Abbreviations

- GAN:** generative adversarial network
- ICD-9:** International Classification of Diseases, Ninth Revision
- ICU:** intensive care unit
- LOS:** length of stay
- LSTM:** long short-term memory network
- MAE:** mean absolute error
- MLM:** masked language modeling
- NLP:** natural language processing
- SMOTE:** Synthetic Minority Over-Sampling Technique
- TeDi-BERT:** temporal distribution matching applied on BERT

Edited by K El Emam, B Malin; submitted 02.06.23; peer-reviewed by A Tomar, D Barra; comments to author 11.10.23; revised version received 31.12.23; accepted 28.07.24; published 02.10.24.

Please cite as:

Agmon S, Singer U, Radinsky K

Leveraging Temporal Trends for Training Contextual Word Embeddings to Address Bias in Biomedical Applications: Development Study

JMIR AI 2024;3:e49546

URL: <https://ai.jmir.org/2024/1/e49546>

doi: [10.2196/49546](https://doi.org/10.2196/49546)

PMID:

©Shunit Agmon, Uriel Singer, Kira Radinsky. Originally published in JMIR AI (<https://ai.jmir.org>), 02.10.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR AI, is properly cited. The complete bibliographic information, a link to the original publication on <https://www.ai.jmir.org/>, as well as this copyright and license information must be included.

Original Paper

Understanding the Long Haulers of COVID-19: Mixed Methods Analysis of YouTube Content

Alexis Jordan¹, MS; Albert Park¹, PhD

Department of Software and Information Systems, UNC Charlotte, Charlotte, NC, United States

Corresponding Author:

Albert Park, PhD

Department of Software and Information Systems

UNC Charlotte

9201 University City Blvd

Woodward 310H

Charlotte, NC, 28223-0001

United States

Phone: 1 7046878668

Email: al.park@charlotte.edu

Abstract

Background: The COVID-19 pandemic had a devastating global impact. In the United States, there were >98 million COVID-19 cases and >1 million resulting deaths. One consequence of COVID-19 infection has been post-COVID-19 condition (PCC). People with this syndrome, colloquially called *long haulers*, experience symptoms that impact their quality of life. The root cause of PCC and effective treatments remains unknown. Many long haulers have turned to social media for support and guidance.

Objective: In this study, we sought to gain a better understanding of the long hauler experience by investigating what has been discussed and how information about long haulers is perceived on social media. We specifically investigated the following: (1) the range of symptoms that are discussed, (2) the ways in which information about long haulers is perceived, (3) informational and emotional support that is available to long haulers, and (4) discourse between viewers and creators. We selected YouTube as our data source due to its popularity and wide range of audience.

Methods: We systematically gathered data from 3 different types of content creators: medical sources, news sources, and long haulers. To computationally understand the video content and viewers' reactions, we used Biterm, a topic modeling algorithm created specifically for short texts, to analyze snippets of video transcripts and all top-level comments from the comment section. To triangulate our findings about viewers' reactions, we used the Valence Aware Dictionary and Sentiment Reasoner to conduct sentiment analysis on comments from each type of content creator. We grouped the comments into positive and negative categories and generated topics for these groups using Biterm. We then manually grouped resulting topics into broader themes for the purpose of analysis.

Results: We organized the resulting topics into 28 themes across all sources. Examples of medical source transcript themes were *Explanations in layman's terms* and *Biological explanations*. Examples of news source transcript themes were *Negative experiences* and *handling the long haul*. The 2 long hauler transcript themes were *Taking treatments into own hands* and *Changes to daily life*. News sources received a greater share of negative comments. A few themes of these negative comments included *Misinformation and disinformation* and *Issues with the health care system*. Similarly, negative long hauler comments were organized into several themes, including *Disillusionment with the health care system* and *Requiring more visibility*. In contrast, positive medical source comments captured themes such as *Appreciation of helpful content* and *Exchange of helpful information*. In addition to this theme, one positive theme found in long hauler comments was *Community building*.

Conclusions: The results of this study could help public health agencies, policy makers, organizations, and health researchers understand symptomatology and experiences related to PCC. They could also help these agencies develop their communication strategy concerning PCC.

(JMIR AI 2024;3:e54501) doi:[10.2196/54501](https://doi.org/10.2196/54501)

KEYWORDS

long haulers; post-COVID-19 condition; COVID-19; YouTube; topic modeling; natural language processing

Introduction

Background

“It’s like a...like a viral tornado that goes in you and kind of just messes you up,” Sadi Nagamutu says in between labored breaths [1]. This is how the account of the battle with post-COVID-19 condition (PCC) of Sadi Nagamutu, a fitness instructor aged 44 years, began during a news interview [1]. In the comment section of the video, one user wrote the following:

I had to pause the video at 2:20. I broke down in tears because I feel like I’m not alone. I have the same thing.

At the time of recording, Sadi Nagamutu had been a patient with PCC for 8 months. By this time, she claims that it had completely disrupted her life. She notes that she went from being a trainer to not being able to lift grocery bags and walk at the same time [1]. It is clear from the comments left under the 60 minutes video that Sadi Nagamutu is not alone in experiencing a drastic change in her quality of life.

The COVID-19 pandemic has changed the lives of many, though one consequence of it has received less attention [2]. PCC has been identified as a syndrome affecting patients long after their initial COVID-19 infection has cleared. These patients are colloquially called *long haulers* [3]. High ratios of those who have been infected with COVID-19 have persisting symptoms that last months after the initial infection.

Studies have shown that PCC has real implications in people’s everyday lives. The World Health Organization Quality of Life Brief Version, a quality of life questionnaire, was administered among patients who had been hospitalized with COVID-19 [4]. The results showed that 30.2% of respondents had PCC, which affected nearly all domains of quality of life as outlined in the World Health Organization Quality of Life Brief Version criteria [4]. Moreover, there have been recent links between PCC and deteriorating mental health [5].

PCC has negative economic implications as well. Those with PCC are often not in condition to work and, thus, realize their full earning potential [2]. Approximately 44% of people with PCC are completely out of the workplace, whereas 51% have reduced hours at work [2]. This could result in >US \$50 billion in lost income annually [2].

In addition, some patients do not receive insurance coverage for PCC-related testing and treatments [6]. This has led to significant debt for some patients [6]. In May 2023, the *Journal of the American Medical Association* estimated that average PCC-related medical costs could be approximately US \$9000 a year [6]. There is also the issue of lost wages due to PCC, which further complicates the medical debt. Making a case for those with PCC by uncovering patient experiences could be useful for public health officials and medical insurance companies, who may need additional help in understanding how debilitating PCC can be.

Social media is a rich source of information regarding people’s experiences and attitudes [7,8] due to the pervasiveness of social media apps and the freedom with which people engage in

discourse on various topics. Such pervasiveness contributes to the increased size of health-related data [7]. This has encouraged researchers to use computational models to analyze social media texts concerning COVID-19 [7,9-12]. One popular method of analysis is topic modeling. Topic modeling allows for the discovery of thematic relationships and patterns within a body of text using natural language models [10]. Latent Dirichlet allocation (LDA) is a probabilistic unsupervised classification method [13]. It has been widely used in studies using topic modeling on a large set of documents [13].

For example, Mutanga and Abayomi [10] used an LDA topic model to study COVID-19-related posts in South Africa and found that conversations revolved around *alcohol consumption, staying at home, vaccine conspiracy theories, police brutality, statistics training, and 5G* [10]. In addition, other authors have explored public sentiment and discourse on COVID-19 vaccines on Reddit using an LDA model [9]. They found that posts covered the broader discussions of *vaccines, safety concerns, efficacy, and side effects*.

To date, there has been one other study that examined the experiences of long haulers on YouTube in the hopes of understanding web-based health communication [7]. However, Jacques et al [7] did not use any topic modeling methods. Instead, they manually coded the 100 most viewed PCC videos based on a predetermined list of themes.

In the following sections, we provide a review of long haulers and health discussions on YouTube. On the basis of this, we sought to understand what types of videos are available on YouTube regarding long haulers and how users respond to PCC-related content. Next, the procedure of data collection and analysis is provided. Results are then reported regarding salient themes for each type of content creator and positive and negative comments. Finally, we conclude with a discussion of the theoretical and practical implications of our findings.

Related Work

Long Haulers

Studies have focused on long haulers in the hopes of understanding their symptoms and concerns [14-16]. By analyzing Reddit posts, Thompson et al [15] found that discussions revolved around *symptoms, diagnostic concerns, broad health concerns, chronicity, support, identity, and anxiety*. In the study by Basch et al [17], news articles and videos were selected from a news media platform (Google News). They were then analyzed to identify common symptoms that appeared in PCC-related content [17]. The authors found that 41% of news reports mentioned the *duration* of the symptoms, which tended to range between 1 month and over a year. *Tiredness* and *fatigue* were the most mentioned symptoms, occurring in 74% of the news content. Though insightful, these studies do not focus solely on the YouTube platform, wherein there can be interaction between the creators of long-form content and those consuming their media.

Health on YouTube

YouTube is a platform that motivates users to create, publish, and comment on posts [18]. It has been developed to handle

long-form content. YouTube is unique in that creators of long-form content can not only share their videos but also engage with viewers within the comment section. A report from Statista estimates that, as of April 2022, YouTube has 247 million users in the United States [19]. There have been studies in which researchers analyzed YouTube comments and transcripts to understand public sentiment on health-related matters. These studies have used either manual [20,21] or natural language processing–based [22,23] approaches.

Noncomputational analyses of YouTube videos have involved manually coding videos into various groupings. One study on anorexia-related YouTube videos used the help of 3 physicians to categorize 140 videos against a predetermined list of classification criteria [20]. In addition, to understand discourse on YouTube videos that seeks to stigmatize mental health, McLellan et al [21] manually coded 100 randomly selected comments from 20 videos based on predetermined coding criteria.

In contrast, Aslam et al [22] used computational methods to understand the transcripts of 1000 COVID-19–related YouTube videos [16]. They used the Gensim LDA topic model to understand the transcripts. They found that salient topics involved *symptoms*, *precautions*, and *homeremedies* [22]. In their study, Serrano et al [23] fine-tuned the Robustly Optimized Bidirectional Encoder Representations Approach base to label comments from factual and misinformative COVID-19–related videos. In addition, they extracted features from video titles and comments [23]. These features were used in a linear support vector machine to detect misinformative videos [23].

We collected YouTube transcripts and comments between August 3, 2020, and October 29, 2021, to investigate PCC symptomatology and other related complications. We chose to use computational methods, more specifically topic modeling, because they can capture a wider distribution than manual studies [10]. To the best of our knowledge, this is the first study to examine YouTube video transcripts and comments related to PCC experiences. Our research questions (RQs) for this study were two-fold: (1) What types of videos are available on YouTube regarding PCC? (RQ 1) (2) How do users respond to PCC content? (RQ 2).

Methods

Data Collection

YouTube is a free-to-use social media platform that has been adopted by individuals, organizations, and specialized professionals from various fields to share relevant and important information [18]. Because of this, we deemed YouTube to be a good source of data to capture videos uploaded by different types of content creators. This allowed for diversity in our data set.

We used Google’s application programming interface, `googleapiclient.discovery`, to capture video comments and metadata (eg, number of comment likes and responses). Data from the top 50 videos as a result of searching each of the following terms were collected: “Covid Long Haulers,” “COVID-19 Long Haulers,” “Long Covid,” “Long Haul Covid,”

“PASC Covid,” “Post-Covid Symptoms,” and “Post-Covid Syndrome.”

The search terms were found by first inspecting COVID-19 long hauler–related news articles to find pertinent keywords. After this, Google Trends was inspected to see whether there were any additional terms or versions of terms that had already been identified. We used these terms to find and inspect an initial list of videos on YouTube. After completing this process, we were able to rule out the term “Longhauler” as many references were not related to COVID-19. The resulting videos were in the date range between August 3, 2020, and October 29, 2021. The videos were collected on November 1, 2021. After removal of duplicates and irrelevant videos, we collected 152 videos.

We used the Python package *YouTubeTranscriptAPI* (Python Software Foundation) to capture transcripts from the videos. It should be noted that the comments collected in our data-gathering process only reflect the top-level comments. In essence, this means that any replies to the original comments were not captured.

We then manually grouped the videos based on the video source as previously done in a similar study [24]. This is because the topic coverage of videos can vary widely depending on the source of the video. The resulting groups were news sources, medical sources, and long haulers. News source videos were those that were uploaded by news entities, including local, national, and international news stations. News source videos represented 51.3% (78/152) of the collected videos. Medical source videos were those that were posted by medical experts such as physicians, health insurance companies, and medical schools. We collected 49 such videos. The last 16.4% (25/152) of the videos belonged to the long hauler grouping, which represented first-person accounts from those who considered themselves to have PCC. From these videos, we collected 2845 comments in total: 1258 (44.22%) associated with medical source videos, 1078 (37.89%) from news source videos, and 509 (17.89%) from long hauler videos.

Ethical Considerations

We only analyzed publicly available documents in this study and did not analyze identifiable private information or involve any direct or indirect interactions with individuals. Per (blank for review) policy (citation: 45 Code of Federal Regulations 46 Definitions), this study is exempt from institutional review board requirements because it does not meet the regulatory definitions of human participant research. However, we removed any user identifiable information (eg, usernames) and paraphrased or modified comments to preserve user pseudonymity while maintaining the content’s integrity in the manuscript.

RQ 1 Methods: What Types of Videos Are Available on YouTube Regarding PCC?

To understand themes within the video content, we used Biterm to generate topics of video transcripts as well. Biterm topic model learns topics by modeling the generation of word co-occurrence patterns in whole documents to counter the sparse word co-occurrence pattern problem that occurs when evaluating at the document level [25]. Each group—medical sources, news

sources, and long haulers—was processed individually to preserve our groupings. Biterm was created with shorter social media texts in mind given that they are usually much shorter than standard document sizes [25]. Because video transcripts were considerably longer and, thus, could contain multiple topics, chronological batches of 50 consecutive words were fed into each model as suggested by previous work on topic modeling [26]. It was important to divide the transcripts into shorter portions so that more specific topics would be generated. After preprocessing our data by lemmatizing words and removing stop words, we fed our data into the topic models. To preserve our groupings, we created 6 separate models: one positive and one negative model for each group (news sources, medical sources, and long haulers). When fine tuning the number of topics, we tested 4 numbers (3, 5, 10, and 15). For each number, we assessed the coherence scores and strength for words within the same topic co-occurring in the same documents [25]. Biterm adopted a coherence score proposed by Mimno et al [27]. In the study by Yan et al [25], the average coherence score for a Biterm model with 5 topics was between -52.3 and -52.5 . A limitation of the coherence score is that it only accounts for the most frequent topic words. To compensate for this limitation, we complemented the evaluation with manual analysis in addition to considering coherence scores for selecting the most cohesive model. To elicit unknown, emerging themes grounded in the labeled topics, we further qualitatively analyzed comments or transcripts within each topic following an open coding procedure [28] similar to that in a previous study that analyzed social media content that included YouTube videos on COVID-19 [29]. Following the collaborative identification of a list of topic labels, the research team independently labeled each topic using up to 50 most salient terms and up to 30 samples of the most representative content followed by grouping the topics into themes. At each iteration, the research team resolved any discrepancies through discussion.

RQ 2 Methods: How Do Users Respond to PCC Content?

We conducted sentiment analysis to understand public sentiment with regard to the delivered content. We used the Valence Aware Dictionary and Sentiment Reasoner (VADER) [30] to determine the sentiment of video comments. VADER is a rule-based model for sentiment analysis. It was created specifically for social media contexts as it can recognize slang and emojis. It produces positive, negative, neutral, and compound scores for each body of text by summing the valence scores of each word and normalizing them to be between -1 and 1 . We chose VADER in lieu of other sentiment analysis tools such as AFINN, BING, or National Resource Council because VADER was specifically developed for analyzing social media texts.

We used the compound score as the overall sentiment score for the comment. Positive comments included all comments with a compound score of >0 . Negative comments included all comments with a score of ≤ 0 following methodological guidance from a previous study [31]. This process was completed independently for each group (medical sources, news sources, and long haulers).

After we created positive and negative subgroups of the comments, we created topic models to understand the thematic make-up of positive and negative comments with regard to each group. We separated comments into positive and negative subgroupings before generating topics so that our resulting topic models would be more cohesive. Similar to the methods for RQ 1, we used Biterm to generate topics and manual review to label topics and group them according to themes. Because comments are relatively short in length and typically have 1 topic, we used the entire comment as a document.

Results

Overview

We organized the resulting topics into 28 themes across all sources. Medical source transcript themes were *Explanations in layman's terms*, *Show housekeeping*, and *Biological explanations*. News source transcript themes were *Sharing patient experiences*, *Negative experiences*, *Experts weighing in*, and *Handling the long haul*. Long hauler transcript themes were *Taking treatments into own hands*, *Changes to daily life*, and *Choosing homeopathy over pharmaceuticals*. Positive news source comment themes were *Extending empathy*, *Expressing distrust through sarcasm*, and *Encouragement for better outcomes*. News source videos received the highest proportion of negative comments. Negative news source comment themes were *Reproduction of debunked and political theories*, *Misinformation and disinformation*, and *Issues with the health care system*. In contrast, medical source videos received the highest proportion of positive comments. Positive medical source comment themes were *Appreciation of helpful content*, *Hope and encouragement*, and *Exchange of helpful information*. Negative medical source comment themes were *Negative impacts of long haul*, *Requiring medical alternatives*, and *Lack of needs*. Positive long hauler comment themes were *Appreciation*, *Exchange of helpful information*, and *Community building*. Negative long hauler comment themes were *Exchange of additional information*, *Disillusionment with the health care system*, and *Requiring more visibility*.

RQ 1 Results: What Types of Videos Are Available on YouTube Regarding PCC?

Overview

We collected the transcripts from 152 videos that were divided into 3 groups (news sources: $n=78$, 51.3%; medical sources: $n=49$, 32.2%; and long haulers: $n=25$, 16.4%). Transcripts were divided into subgroups of 50 consecutive words and fed into distinct Biterm topic models. The following sections show the breakdown of videos by source type.

Medical Source Video Transcripts

Overview

The medical source video transcripts were captions from videos created by an individual or organization in the medical sector. This included physicians, medical insurance companies, and medical schools.

Explanations in Layman's Terms

The first theme, *Explanations in layman's terms*, covered 3 topics: "Symptomatology," "Symptom etiology," and "Symptom management." As implied by the theme title, the transcript snippets constituting each topic displayed scientific speech that was relatively easy for the public to understand. The first topic, *Symptomatology*, covered video transcripts in which the speaker explained the symptoms associated with COVID-19. Some medical source content creators dedicated entire videos to just a few symptoms or a particular health system, as was the case in a video from University of Alabama at Birmingham medicine dedicated to PCC and hair loss:

...when you go through something stressful and you have a telogen effluvium, most of your hairs can enter the resting phase at the same time.

The second topic, "Symptom etiology," featured transcript snippets that offered explanations of how PCC symptoms might have originated. Finally, "Symptom management" featured transcript snippets wherein medical professionals offered potential treatments for symptoms.

Show Housekeeping

Show housekeeping was another prevalent theme in medical source video transcripts. Associated topics were "Introducing the show or guest," "Validating guests' credentials as a reliable source," and "Encouraging the audience to keep in touch." As the name suggests, these videos routinely introduced each of the medical experts on the show and expounded on their

credentials. This could potentially be due to the idea that many information consumers can be critical of the source of their information. Expounding on the guest speakers' credentials could help build credibility and trust between the video publisher and the audience. The next topic dealt with encouraging the audience to keep in touch. Some medical source content creators offered links to other social media platforms where they could continue the long hauler conversation with engagers.

Biological Explanations

In general, biological explanations comprised transcript snippets that displayed more advanced scientific language than that shown in the *Explanations in layman's terms* theme. Biological explanations featured 2 distinct topics: "Immunophenotyping" and "Explaining the mechanics of immune responses." Immunophenotyping is the process of identifying cells based on antigens or markers [32]. In one video, the speakers discussed using "proprietary spark dyes," which can be used for immunophenotyping [32]. In addition, these videos were concerned with explaining the mechanics of immune responses. In this case, a biological perspective of disease etiology was offered, with less use of layperson terminology.

News Source Video Transcripts

Overview

The news source video transcripts were the captions from news media outlets. These outlets ranged from local to international audiences (Table 1).

Table 1. News source video transcript results.

Theme and topic label	Keywords	Sample transcripts
Sharing patient experiences		
Symptoms	“Patients,” “symptoms,” “life,” “understand,” “hair,” “feel,” “medical,” “sick,” “protein,” “heart-beat,” “health,” and “doctors”	“Five months later, she is still short of breath. Doing therapy three times a week. It often feels like this body is not mine. That the things that i want to do i can’t do.”
Treatments	“Need,” “better,” “understand,” “doctors,” “months,” “trying,” “research,” “care,” and “answers”	“[...] even though there’s not a magic pill yet, to cure a long COVID, at least we can try to aggressively manage the symptoms, connect them with other patients, other resources, and try to help in whatever way we can.”
Negative experiences		
Not being believed by others and doctors	“Symptoms,” “covid,” “virus,” “physician,” “dr,” “feeling,” “need,” “progress,” and “says”	“[...] to those doctors that deny the existence of long covid that this thing of course it’s really look at the science.”
Explaining the impact of “long Covid” on lives	“Started,” “need,” “progress,” “end,” “taken,” “coming,” “time,” “medical,” “virus,” “smell,” “health,” “watch,” and “feeling”	“Differently, less like the flu and more like a condition that can have lasting repercussions. The moment [...] the sick get to go home. But for many it’s not the end, it’s just the beginning of a long and perilous road to recovery.”
Experts weighing in		
Etiology of the disease	“Effects,” “infection,” “different,” “virus,” “actually,” “research,” “seen,” “syndrome,” “persistent,” and “fatigue”	“Today chris hrapsky talked with an expert whose theory on this is gaining attention. Mast cells are the first responders of your immune system when an infection occurs in under a second these cells and stuff like histamine to other cells to say, hey, wake up, something’s wrong here. In some people these mass cells go hay-wire and overreact like central dispatch calling in the swat team for a coffee spill at starbucks and this is called mast cell activation syndrome.”
Experts explaining “long Covid”	“Struggle,” “lingering,” “illness,” “health,” “syndrome,” “persistent,” “group,” “body,” “covid19,” “physical,” and “related”	“The other thing that makes it really challenging, is that symptoms are not necessarily always correlated or equal to organ dysfunction that we can measure [...].”
Handling the long haul		
Managing symptoms	“Test,” “hair,” “brain,” “doctor,” “fatigue,” “pain,” “disease,” “talk,” “common,” “exercise,” “need,” “home,” and “levels”	“[...] they sent an occupational therapist to see what they could do in the house so our washroom has been retrofitted with a brand new high toilet because he had issues getting on and off the toilet [...].”
Handling cardiac or chest problems specifically	“Oxygen,” “need,” “lung,” “blood,” “chest,” “pain,” “infection,” “shortness,” “ventilator,” “loss,” “pulmonary,” “complications,” “attack,” “disease,” and “breath”	“[...] what i suggest is that those of our patients who are having tachycardia it’s not a bad idea to get themselves screened by their physicians or cardiologists so that at least we are clear that a patient does not have baseline pulmonary embolism [...].”

Sharing Patient Experiences

“Symptoms” and “Treatments” are 2 topics that were part of the *Sharing patient experiences* theme. The *Symptoms*-related video transcripts dealt with interviewees sharing their daily symptoms to give perspective to audiences. Interviewees experienced a wide range of symptoms. These symptoms appeared to have a significant impact on daily life. One interviewee noted that she would fall due to elevated heart rate that worsened doing routine tasks such as “just walking from here to the kitchen.” Guests were also concerned with finding some type of treatment that could mitigate PCC symptoms. Patients seemed to have managed expectations regarding treatment but exhibited some level of hope:

...there’s not a magic pill yet, to cure long Covid...at least we can try to aggressively manage the symptoms.

Negative Experiences

The *Negative experiences* theme featured 2 related topics. The first topic was “Not being believed by others and doctors.” This was a particularly common topic throughout the text. Interviewees shared their experiences of being ignored or not believed. These long haulers sought and could not find affirmation:

...no one really understands me.

The next topic dealt with explaining the impact of PCC on lives. Long haulers and news reporters introduced PCC in general terms as well as the people that it had impacted. One long hauler explained the following:

...nearly seven months later and I’m still unwell and I am still a broken woman.

Experts Weighing In

The *Experts weighing in* theme had 2 topics: “Etiology of the disease” and “Experts explaining long Covid.” Similar to medical source videos, experts took 2 approaches when speaking about PCC. The first approach, as evidenced in the *Etiology of the disease* topic, explained things from a strictly biological perspective:

Mast cells are the first responders of your immune system when an infection occurs.

In contrast, in *Experts explaining long Covid*, more commonly used colloquial language was used to explain PCC:

...different studies use different thresholds, which makes it really challenging to compare apples to apples.

Handling the Long Haul

The last theme had 2 topics as well: “Managing symptoms” and “Handling cardiac or chest problems specifically.” *Managing symptoms* dealt mainly with long haulers finding their own ways to manage their illness. In addition, cardiac and chest problems were often discussed. They are common symptoms that were addressed by experts and patients alike. Experts offered symptom management advice:

...and it will take three to six months for this myocarditis to settle.

Long Hauler Video Transcripts

Overview

The long hauler video transcripts were captions from individual content creators that talked directly about their own personal experiences with PCC (Table 2).

Table 2. Long hauler video transcript results.

Theme and topic label	Keywords	Selected sample transcripts
Taking treatment into own hands		
Alternate remedies	“New,” “health,” “try,” “better,” “care,” “fungus,” “changing,” “trusted,” and “broken”	“Covid was my wake-up call to fix my gut and ultimately fix my health I was declining I was already declining before covid I was getting weak [...]”
Dealing with uncertainty	“Changing,” “declining,” “shitty,” “new,” “life,” and “work”	“[...] this is my story right like this is this is what I have to live with for an indefinite period of time so my very good family friend she runs her own practice she’s an MD and she said you know like nobody should want to get Covid because nobody knows the lasting effects of Covid.”
Not being listened to by physicians	“Biases,” “trusted,” “chore,” “dr,” “doctors,” “feel,” “medicine,” and “care”	“[...] especially female patients and patients of color the benefit of the doubt [...] there is so much research on patients reporting doctors not believing them or not treating them with the same level of compassion [...] I didn’t think it would happen to me [...]”
Changes to daily life		
Insomnia	“Helped,” “started,” “pills,” “prevent,” “restless,” “waking,” and “blockers”	“I am allowed to take a maximum amount of the sleeping aids and they don’t work I just get a calming feeling along with my multitude of symptoms I think along with the drenching sweats and the fevers that just won’t stop because my husband has to cover me in ice sometimes because even with medication the fever doesn’t stop climbing.”
How symptoms interrupt activities	“Day,” “symptoms,” “time,” “feel,” “bad,” “need,” “breath,” “overgrowth,” “taste,” “chronic,” “fever,” “life,” “nuts,” and “sacrifice”	“I had to stop eating eggs I recognize that eggs weren’t agreeing with me anymore and [...] I was eating three eggs every day like that was you know that was a breakfast staple for me [...]”
How symptoms present themselves	“Experience,” “fever,” “health,” “day,” “highly,” “discovering,” “seizures,” “entry,” and “permanent”	“So like whenever i would get near like the oven or the stove or like the air fryer or take a shower or try to exercise like whenever my internal body temperature would rise my face would go bright red it would get swollen id’ get like weird patches it was super strange [...]”
Choosing homeopathy over pharmaceuticals		
Use of CBD ^a and THC ^b	“Gummies,” “high,” “work,” “try,” “started,” “need,” and “help”	“The cbd and the gummies that I take to sleep at night [...] I just try to keep things as natural.”
Turning down over-the-counter medicine	“Deficiency,” “vitamin,” “blood,” “different,” “taking,” and “bad”	“[...] so naturally I assume that is still coronavirus so she encouraged me to take over the counter medication which I don’t do i’ve never done it I don’t do it I don’t believe in it I don’t have a Tylenol deficiency I don’t have an aspirin deficiency i’m not ibuprofen deficient so I don’t think I should take that.”

^aCBD: cannabidiol.

^bTHC: tetrahydrocannabinol.

Taking Ownership of Treatment

The 3 related topics were “Alternate remedies,” “Dealing with uncertainties,” and “Not being listened to by physicians.” *Alternate remedies* dealt with long haulers sharing alternative medicine that they used and recommending alternative medicine to others. In *Dealing with uncertainties*, long haulers noted that they were dealing with symptoms for “an indefinite period of time.” On the basis of their experiences, they had an understanding that physicians were mystified by PCC and, thus, treatments were not certain or foolproof. This led to the last topic, which was “Not being listened to by physicians.” A recurrent topic thus far in the study, this dealt with patients not feeling listened to and supported by members of the health care system. One particularly popular account of this was shared by one woman in a video titled “I’ve had COVID-19 for a year. Here’s what I’ve learned.” She shared her experience as a woman and person of color who felt that she experienced particularly unfair treatment:

...there is so much research on patients reporting doctors not believing them or treating them with the same level of compassion.

Long haulers called for physicians to hold themselves accountable when confronting their own biases. If not, long haulers suggested that they were “violating the trust of their patients and trust is a key element to the patient physician relationship.”

Changes to Daily Life

Next, long haulers discussed the impact of the long haul on daily life. Associated topics included “Insomnia,” “How symptoms interrupt activities,” and “How symptoms present themselves.” Long haulers discussed how insomnia impacted their lives. They mentioned that their symptoms impeded their ability to exercise, eat foods they regularly ate, and even take showers. Finally, long haulers talked about how the symptoms initially presented themselves.

Choosing Homeopathy Over Pharmaceuticals

The 2 related topics were “Use of CBD and THC” for treatment and “Turning down over-the-counter medicine.” One long hauler looked to tetrahydrocannabinol gummies to cure insomnia in part because “I don’t like pharmaceuticals, I have never really

liked them.” Other long haulers shared their apprehension about using pharmaceutical drugs and mentioned turning to more natural options instead.

RQ 2 Results: How Do Users Respond to PCC Content?

Overview

To understand how users respond to PCC content, we separated comments for each category (news sources, medical sources, and long haulers) into 2 subcategories based on sentiment (negative and positive). We then used Bitern to generate topics for these subcategories. When looking at all sources combined, there was not a large discrepancy between the number of positive and negative comments. Overall, there were 1463 positive comments and 1382 negative comments.

However, when we began to look at the split of positive and negative comments by source, we could see that news sources received a greater share of negative comments. There were 687 negative comments and 391 positive comments. In contrast, medical sources received more positive than negative comments. There were 528 negative comments compared to 730 positive comments. Finally, long hauler videos only showed a 13-point difference between the number of positive and negative comments. There were 261 positive comments and 248 negative comments.

In addition to capturing the comments themselves, we captured metadata associated with the comments. This included comment replies, comment likes, and video description. Comment likes and replies indicate the level of engagement that other YouTube users had with the comment posted. Medical source video commenters saw an average of 16.02 (SD 45.09) likes per comment. The most liked comment received 602 likes. The most replied to comment received 474 replies. Conversely, news source video commenters saw an average of 36.46 (SD 168.34) likes per comment. The most liked comment received 2520 likes. The most replied to comment received 184 replies. Finally, long hauler video commenters saw an average of 54.55 (SD 246.72) likes per video. The most liked comment received 4127 likes. The most replied to comment received 223 replies ([Table 3](#)).

Table 3. Comment metadata.

Source	Number of comment likes, mean	Most likes on a comment, N	Number of comment replies, mean	Most replies on a comment, N
Medical source videos	16.02	602	3.15	474
News source videos	36.46	2520	4.23	184
Long hauler videos	54.55	4127	3.06	223

News Source Video Comments

Overview

[Table 4](#) shows the resulting topics and themes from positive comments found under news source videos.

Table 4. Results of positive comments in news source videos.

Theme and topic label	Keywords	Sample comments
Extending empathy		
Relating to others	“People,” “get,” “think,” “symptoms,” “many,” and “felt”	“omg, i can totally relate.”
Well wishes	“Everyone,” “really,” “hope,” “take,” “care,” “able,” and “want”	“It would be terrible to lose your ability to taste or smell. Here’s to hoping they improve soon.”
Gratitude	“Better,” “still,” “hope,” “people,” “feel,” “heart,” “help,” “something,” and “good”	“Your story was gut-wrenching, but still worth the share. Thank you. People need to hear this.”
Expressing distrust through sarcasm		
Sarcasm	“See,” “think,” “often,” “say,” “real,” and “know”	“Okay so they survived a cold like most do. With a 99.8% survival rate, I’m sooo surprised.”
Encouragement for better outcomes		
Prayers and scriptures	“Unto,” “shall,” “ye,” “people,” “peace,” “hath,” “forgive,” “love,” “reward,” “presence,” “pray,” “temple,” and “holy”	“Phillippians 4:7—And the peace of God, which surpasses all comprehension, will guard your hearts and your minds in Christ Jesus.”
Potential solutions and sharing symptoms	“Ask,” “receive,” “keep,” and “know”	“Leronlimab is in clinical trials you guys. Don’t worry, help is on the way.”

Extending Empathy

Extending empathy comprised the topics “Relating to others,” “Well wishes,” and “Gratitude.” Comments in which people related to others involved people explicitly sharing that they related to the content shown or explaining how their symptoms were similar to those of the people interviewed in the news segments. For example, one commenter wrote the following:

You are not alone. I had COVID in April 2020 [...] I am currently in pulmonary rehab [...] I want others to know you are not alone. I’m praying for everyone. God Bless.

Well wishes was the second topic in this theme. In this topic, commenters sought to verbally empathize with those experiencing negative COVID-19–related symptoms:

Too bad for that young man, hopes he gets better!

Finally, in the *Gratitude* topic, commenters were also grateful that PCC content was being shared at all:

So glad she is sharing her struggles.

Expressing Distrust Through Sarcasm

Although the comments observed in this analysis were rated neutral or positive by VADER, some comments seemed to take on a sarcastic tone. For example, one commenter wrote the following:

The greatest nation in the world is your imagoNATION.

These sarcastic comments often appeared to exhibit political or skeptical undertones.

Encouragement for Better Outcomes

The topics within the *Encouragement for better outcomes* theme were “Prayers and scriptures” and “Potential solutions and sharing symptoms.” Many commenters left prayers and extensive Bible verses underneath videos as a form of encouragement for those battling PCC:

God heal these people from this virus. Give them strength.

Finally, *Potential solutions and sharing symptoms* was a topic that covered suggestions that commenters made to improve the symptoms of those dealing with PCC as well as sharing symptomatology in general:

Leronlimab is in clinical trials you guys. Don’t worry, help is on the way.

Negative News Source Comments

Table 5 shows the resulting topics and themes from negative comments found under news source videos.

Table 5. Results of negative comments in news source videos.

Theme and topic label	Keywords	Sample comments
Reproduction of debunked and political theories		
Conspiracy theories	“Vaccine,” “face,” “different,” “affected,” “system,” “situation,” “avoid,” “dreadful,” “resist,” and “corona”	“This is all because of 5G poisoning.”
Political influences	“Capability,” “dreadful,” “never,” “responsible,” “fight,” and “answer”	“They got their butts kicked by Kung flu.”
Misinformation and disinformation		
Fear of impending doom	“Never,” “stress,” “know,” “affected,” “death,” “worry,” and “stop”	“Something’s coming, and we won’t be able to stop it.”
Skepticism or rationalization	“Already,” “must,” “know,” “affected,” “response,” “another,” “nothing,” and “personal”	“Elderly people are susceptible to viruses. This is well known.”
Issues with the health care system		
Not believed	“Medical,” “sick,” “normal,” “pain,” “feeling,” “never,” “anxiety,” “help,” “see,” “hope,” and “think”	“My primary care physician doesn’t believe me either [...].”
Other illnesses	“Sick,” “heart,” “pain,” “long,” “time,” “blood,” “fatigue,” “brain,” “chronic,” and “help”	“I had this for decades with me/cfs. Imagine dealing with it for that long [...].”

Reproduction of Debunked and Political Theories

This theme comprised 2 topics: “Conspiracy theories” and “Political influences.” As an example of the *Conspiracy theories* topic, one commenter offered alternate causes of PCC symptoms, which were based on public disdain for mask wearing—“‘Long-haulers’ may actually be suffering from effects of prolonged mask-wearing [...]”—instead of on veritable information. In contrast, *Political influences* covered suspected country or political involvement that contributed to the pandemic. When referring to individual damages incurred due to PCC, one commenter wrote the following:

...take the cost off the debt to china.

Distrust of Information Shared

This theme comprised 2 topics: “Fear of impending doom” and “Skepticism or rationalization.” *Fear of impending doom* comprised comments that pointed to a grim future for long haulers or the public:

...they’re just trying to kill all the long haulers when all you need is some ivermectin [...]

Skepticism or rationalization comprised commenters who were not convinced that the information presented on PCC was veritable:

...they had flu colds bacterial lung infections pneumonia, many caused by face mask, no sunlight, fear and confinement [...]

Issues With the Health Care System

This theme comprised 2 topics. “Not believed” covered comments condemning health care workers for dismissing the symptoms of their patients:

...typical doctor behavior: when in doubt, blame anxiety.

Other illnesses covered comments in which people drew similarities between PCC and other chronic illnesses:

This is so real...the Lyme community feels all your pain. And being denied by Dr’s that this is real. Its criminal to ignore this.

Medical Source Video Comments

Overview

Table 6 shows the resulting topics and themes from positive comments found under medical source videos.

Table 6. Results of positive comments in medical source videos.

Theme and topic label	Keywords	Sample comments
Appreciation of helpful content		
Gratitude	“Help,” “medical,” “doctors,” “hope,” “thank,” “much,” “understand,” and “positive”	“This is the first thing that I have seen that explains anything besides the news trying to sensationalize and leave out important details.”
Health literacy	“Need,” “information,” “understand,” “research,” “know,” “specific,” and “narrative”	“Your lectures are always easy to understand. Thank you Dr.”
Hope and encouragement		
Prayers	“Hope,” “believe,” “feeling,” and “days”	“Jesus loves you [...]”
Voice of reason	“Help,” “know,” “say,” “think,” “test,” and “research”	“You’ve always cared, been of a sound mind, and shared such insightful information. Thank you.”
Bravery	“Research,” “doctor,” “video,” “help,” “know,” “people,” “feel good,” “positive,” and “believe”	“Even though this subject is controversial, you’re still brave enough to comment on it. Thank you.”
Exchange of helpful information		
Seeking additional information	“Symptoms,” “help,” “information,” “video,” “wonder,” and “please”	“Has the Dr. released the additional information?”
Seeking translated information	“Please” and “videos”	“Can you please translate to Arabic?”
Sharing helpful information	“Think,” “information,” “help,” “vitamin,” “research,” “medical,” and “may”	“Yesterday, I saw an article that said we needed to be aware of [...]”

Appreciation of Helpful Content

This theme covered 2 topics: “Gratitude” and “Health literacy.” *Gratitude* covered general professions of thanks for the content shown. One commenter wrote the following:

Dr. Hansen, this is exactly the information I was hoping for! Thank you.

Health literacy in this case was covered in a positive light. Commenters thanked content makers for presenting information in a clear manner:

...as a lay person with zero medical background, I learn a lot.

Hope and Encouragement

This included 3 topics: “Prayers,” “Voice of reason,” and “Bravery.” *Prayers* included well wishes for those dealing with PCC or reading the comment section. This included requesting prayers as well:

Please pray for my mom...she is positive for covid 19.

The *Voice of reason* topic alluded to the idea that commenters deemed it important to find useful and truthful information:

Thank you for your commitment to keeping the world informed.

Finally, *Bravery* featured comments that alluded to the negativity that those sharing information about PCC and, more generally, COVID-19 face. One commenter noted the following:

...this subject is controversial and you’re still brave enough to comment on it.

Exchange of Helpful Information

This theme covered 3 topics: “Seeking additional information,” “Seeking translated information,” and “Sharing helpful information.” *Seeking additional information* featured those primarily asking questions such as the following: “What about cutaneous hyperesthesia?” In *Seeking translated information*, many sought to understand content by having it translated into their native language. In *Sharing helpful information*, commenters tried to share what they deemed to be helpful to others:

Find a hyperbaric oxygen therapy chamber and a doctor checkup for compassionate use.

Negative Medical Source Comments

[Table 7](#) shows the resulting topics and themes from negative comments found under medical source videos.

Table 7. Results of negative comments in medical source videos.

Theme and topic label	Keywords	Sample comment
Negative impacts of the long haul		
Comorbidity	“Fatigue,” “disease,” “symptoms,” “pain,” “chronic,” “heart,” “brain,” “chest,” “feel,” “body,” and “diagnosed”	“How does this effect those with diabetes. I’m experiencing a range of symptoms.”
Loss	“Family,” “end,” “life,” “months,” and “never”	“COVID-19 took my mom last year. I don’t know how I’ll move on.”
Symptoms	“Symptoms,” “fatigue,” “disease,” “chest,” “heart,” “brain,” “taste,” “body,” “severe,” “hearing,” and “memory”	“I had a headache so bad that I had to seek treatment [...]”
Requiring medical alternatives		
Criticism of physicians	“Doctor,” “bad,” “know,” “experience,” “need,” “study,” “poor,” “data,” and “must”	“These doctors have no idea what they’re doing. His advice makes no sense. I think we’ll be sick forever.”
Debunked recommendations	“Study,” “poor,” “suffering,” “need,” “research,” “must,” and “know”	“How could you share so much but not talk about Ivermectin? You’re doing everyone an injustice.”
Misinformation	“Vaccine,” “last,” “illness,” “death,” and “bad”	“Misinformation has gotten so bad that my own family won’t even believe me [...]”
Lack of needs		
Lack of improvement	“Symptoms,” “feel,” “since,” “weeks,” “pain,” “back,” “effects,” “suffering,” “last,” and “vaccination”	“The vaccine didn’t improve my symptoms.”
Lack of information	“Medical,” “would,” “think,” “cause,” “whether,” and “help”	“He mentions promising treatments but he never tells us what they are.”

Negative Impacts of the Long Haul

This theme comprised 3 topics: “Comorbidity,” “Loss,” and “Symptoms.” *Comorbidity* featured comments and questions that sought to relate PCC to other diseases:

Childhood obesity might be a factor [...]

In *Loss*, some commenters spoke explicitly about those they lost to PCC or COVID-19. Finally, in *Symptoms*, commenters spoke candidly about the symptoms they faced:

I had a headache so bad that I had to seek treatment.

Requiring Medical Alternatives

In this theme, there were 3 topics: “Criticism of physicians,” “Debunked recommendations,” and “Misinformation.” In *Criticism of physicians*, commenters spoke about how they often felt dismissed by physicians when presenting their symptoms:

...if he went to visit my gp he would tell him he was stressed and it was in his head told me the same [...] it turned out to be lung scarring and a tumor.

In *Debunked recommendations*, commenters pushed for the use of medications that had already been proven to be not helpful and even toxic for human consumption. Ivermectin was notably one of these medications:

...should we be taking Ivermectin since our DNA now expresses spike protein forever?

Finally, *Misinformation* comments reverberated common antimask and antivaccine comments:

John do you have the list of ingredients of the vaccines? My daughter makes cupcakes and she has to list every ingredient by law...

Lack of Needs

This theme covered “Lack of improvement” and “Lack of information.” *Lack of improvement* largely related to symptoms not improving despite medical and home remedy attempts. *Lack of information* included criticism of content sources for not providing enough information regarding content (eg, treatment and research):

...he mentions promising treatments, but he never tells us what they are.

Long Hauler Video Comments

Overview

Table 8 shows the resulting topics and themes from positive comments found under long hauler videos.

Table 8. Results of positive comments in long hauler source videos.

Theme and topic label	Keywords	Sample comments
Appreciation		
Bravery	“Sharing,” “thank,” “glad,” “feeling,” “believe,” “recovery,” “care,” “story,” “bless,” and “post”	“Your bravery hasn’t gone unnoticed. Thank you for all that you do.”
Compliments	“Bless,” “share,” “thank,” “good,” “feel,” “positive,” “love,” and “believe”	“what a beautiful person inside and out.”
Exchange of helpful information		
Seeking additional information	“Back,” “still,” “help,” “know,” “could,” “think,” “test,” “say,” “check,” “treatment,” “better,” “work,” “natural,” “support,” and “right”	“...eliminating carbs could potentially make things better. That’s what worked for me [...].”
Sharing additional information	“Support,” “scheduling,” “doctor,” “right,” “treatment,” “better,” “work,” “great,” “different,” “try,” “may,” and “take”	“If you check my channel you’ll see why you should check your CRP. It could really help your lungs [...].”
Community building		
Reaching out	“Need,” “people,” “help,” “sharing,” “video,” “time,” “want,” “research,” “appreciate,” and “experience”	“...is there any way that I can talk to you please or message you?”

Appreciation

The *Appreciation* theme comprised “Bravery” and general “Compliments.” Commenters lauded the content creator for being brave enough to share their experiences. This may allude to the idea that some who speak on their PCC experiences may face backlash. In addition, commenters gave content creators various accolades regarding their personalities and their decisions to share information:

...what a beautiful person, inside and out.

Exchange of Helpful Information

“Seeking additional information” and “Sharing additional information” were the 2 topics in this theme. Commenters often

initiated or tried to engage in dialogue about topics such as potential treatments and tests for PCC symptoms:

...if you check my channel, you’ll see why you need to check your CRP.

Community Building

“Reaching out” was the topic in this theme. Commenters sought to connect with long haulers to continue conversations elsewhere.

Negative Long Hauler Source Comments

Table 9 shows the resulting topics and themes from negative comments found under long hauler videos.

Table 9. Results of negative comments in long hauler source videos.

Theme and topic label	Keywords	Sample comments
Exchange of additional information		
Asking follow-up questions	“Get,” “think,” “covid,” “would,” “take,” “symptoms,” “different,” “know,” “since,” “less,” “help,” and “maybe”	“Are you still sick?”
Sharing information via experience	“Symptoms,” “many,” “swollen,” “frustrating,” “lymph,” “body,” “covid,” “chest,” “get,” “pain,” “take,” and “help”	“You should check into your thyroid levels. I had issues with mine [...].”
Seeking answers for symptoms	“Help,” “feel,” “usually,” “maybe,” “less,” and “symptoms”	“did anyone else experience long COVID anxiety?”
Disillusionment with the health care system		
Disappointment with physicians	“Medical,” “pain,” “enough,” “nurse,” “support,” “right,” “hard,” “felt,” “know,” “fear,” “frustrating,” “people,” and “deal”	“Overachievers will never admit they don’t know something.”
Unfair treatment	“People,” “frustrating,” “deal,” “problem,” “felt,” “hard,” and “know”	“...women and women of color are often treated this way. i’m not really surprised.”
Requiring more visibility		
Gratitude	“Symptoms,” “sick,” “since,” “without,” “feeling,” “believe,” and “almost”	“My girlfriend has long COVID and she has so many of these issues.”
Wanting more awareness	“Weeks,” “always,” “sick,” “bad,” “body,” “infection,” “low,” “feel,” “know,” and “symptoms”	“I barely see any information like this in the media. Why is that?!”

Exchange of Additional Information

“Asking follow-up questions,” “Sharing information via experience,” and “Seeking answers for symptoms” were the 3 topics in this theme. This theme was very similar to the theme that appeared in the positive long hauler comment analysis. There were slight differences between the examples in the 2 themes. The theme in this instance focused more on symptomatology in the case of the content creators or commenters:

...did anyone else experience long COVID anxiety?

Disillusionment With the Health Care System

The topics in this theme were “Disappointment with physicians” and “Unfair treatment.” In “Disappointment with physicians,” commenters mainly criticized the behavior of physicians in the context of PCC diagnosis or lack thereof. In addition, in “Unfair treatment,” commenters mentioned how specific groups may experience worse health care treatment than others:

...female patients and patients of color [...] there is so much research on patients reporting doctors not believing them or not treating them with the same level of compassion.

Requiring More Visibility

This theme comprised “Gratitude” and “Wanting more awareness.” Interestingly, although these comments were marked as negative, there were still a number of comments that expressed gratitude for the content creator sharing their message. This was often accompanied by sharing of their experiences as well. Relatedly, “Wanting more awareness” reflected the desire of commenters to see additional PCC content in the media, insinuating that there was not yet enough coverage.

Discussion

Principal Findings

Overview

Symptomatology was a prevalent theme across all sources. Video creators and commenters shared and empathized with each other regarding symptoms that occurred because of PCC. These symptoms included *prolonged fatigue, cognitive dysfunction, shortness of breath, cardiac issues, and lingering pulmonary symptoms*. This was consistent with the findings of several studies [4,14-16,33]. In medical source videos, medical professionals explained symptomatology and symptom etiology in both layperson and more scientific terms. In both news source and long hauler videos, personal experiences were shared, as well as how PCC symptoms had impacted their daily lives. Upon inspection of the comments, we found that symptoms were shared for a range of purposes. At times, it was purely to exchange knowledge and offer informational support. In addition, it was used as a means to connect with others to exchange emotional support [34].

Emotional and Informational Support

The positive themes identified in our findings can be operationalized as emotional and informational support. The emotional support category of themes comprised those in which

commenters or video creators sought to empathize with others. This was through words of encouragement, prayers, sharing of similar experiences, and community building. Informational support themes covered themes in which users sought or shared information.

In both transcripts and comments, people discussed experiences of not being believed by physicians and having a perilous relationship with the health care system. This sentiment appeared to be common across the board; however, 3 groups stood out in particular: those with other chronic illnesses such as chronic fatigue syndrome and myalgic encephalomyelitis, women, and people of color. Those who had been battling chronic diseases for years before the emergence of PCC empathized with long haulers who felt that they were not being heard, as can be seen in Table 5. Complaints centered on being told that they were overexaggerating their symptoms or insinuations that patients were hypochondriacs (Table 5).

Women and people of color discussed how they felt dismissed by health care workers. There was a general sentiment of distrust. This notion has been backed by an NBC article, wherein one woman of color explained that she had been brushed off by physicians and labeled as aggressive [35]. This was despite the fact that she had lost 30 pounds and sight in her right eye as a result of PCC [35]. People of color have been disproportionately affected by PCC [35-37]. A total of 2 studies conducted by the National Institutes of Health [36,37] found that Hispanic and African American individuals had greater health problems and symptoms related to PCC but were less likely to be diagnosed. This corroborates anecdotal evidence from video comments (Table 9).

Though these topics occurred in comments with negative sentiment, there were positive repercussions: emotional and informational support. This general distrust of the health care system appears to have led to the adoption of homeopathic medicine, alternative medicine, and home remedies. In attempts to take their health into their own hands, users resorted to alternative treatments even if it put their freedom at risk. These comments were shared freely between video creators and commenters, exemplifying informational support. For example, one commenter noted that they smuggled marijuana into their state and felt that their insomnia had improved as a result of consuming it (Table 8). Others suggested changes in dietary habits (Table 8).

Another aspect of informational support dealt with health literacy. Health literacy was a theme that appeared most often in medical source-related videos. Health literacy has been defined by the Centers for Disease Control and Prevention as the degree to which individuals have the ability to understand and use information to make health-related decisions [38]. The content from medical sources exhibited 2 distinct tones. In the first case, information was delivered in layperson terms, which would likely be easier for the average person to understand. In the second case, scientists presented biological explanations of PCC in more jargon-filled language. There were mixed reactions. Commenters noted that, at times, they had issues understanding the content (Table 7). Issues with health literacy can impede one's ability to properly advocate for themselves

and understand what their options are. In other instances, commenters thanked the medical professionals for explaining PCC in a digestible manner, as can be seen in [Table 6](#).

Symptom management was another topic that came up often in medical source and long hauler video transcripts and comments.

In videos, medical professionals outlined steps that those with PCC could take to mitigate their symptoms ([Table 10](#)). In the comment section of medical source videos, commenters shared helpful information as well ([Table 6](#)).

Table 10. Medical source video transcript results.

Theme and topic label	Keywords	Sample transcripts
Explanations in layman's terms		
Symptomatology	"Symptoms," "long," "fatigue," "common," "brain," "pain," "loss," "breath," "chest," "shortness," "smell," "body," "fog," "taste," "breathing," and "cough"	"Cognitive impairments things like word finding difficulty, short-term memory loss, difficulty with multitasking, poor concentration as well as anxiety and PTSD especially in patients who have been hospitalized."
Symptom etiology	"Syndrome," "severe," "illness," "chronic," and "different"	"[...] As I said earlier all of these symptoms, the headache, the sleep disturbance, the brain fog, they often tend to run together and sometimes it's hard to say as to what is leading to what other symptom. It's sort of like the chicken and the egg analogy. Is it because somebody has poor sleep, is that what leads to headaches because we do know what headaches can be triggered when the sleep is poor."
Symptom management	"Vitamin," "time," "day," "sleep," "work," "different," "need," "help," and "right"	"I think the first treatment for that insomnia is really sleep hygiene so that's things like um turning off devices a half an hour before bed time, making sure you go to bed at the same time with a relaxing bedtime ritual, waking up at the same time every day, shutting devices off. [...]"
Show housekeeping		
Introducing the show or guest and validating the guest's credentials as a reliable source	"Dr," "going," "thank," "time," "want," "talk," "need," "help," "work," "research," "medical," and "information"	"And he has organized several conferences given many lectures and has done live surgeries as demonstrations in several international conferences and forums [...] and nhs hospitals in UK [...]"
Encouraging the audience to keep in touch	"Question," "talk," "help," "data," "group," "better," and "information"	"[...] There's a conversation on X hashtag covered science um and uh all that remains then is for me to thank everyone that's submitted questions. I hope I got through as many as I could."
Biological explanations		
Immunophenotyping	"ccr5," "data," "number," "antigen," "cd16," "cd14," "interleukin," "dotted," "chord," "monocytic," and "interstitial"	"[...] You know we can apply these variants in multiple applications, such as immunophenotyping cell sorting and also to study cell physiology [...]"
Explaining the mechanics of immune responses	"Disease," "percent," "inflammation," "severe," "heart," "illness," "brain," "course," "study," and "viral"	"[...] What's really interesting is interleukin-2 and interferon-gamma are two cytokines that are intimately involved in antiviral immune responses and they are low in active because it's an emerging infection our immune system presumably has not seen that virus before [...] so long covid actually has an immune response with high interferon-gamma that looks very much like a typical antiviral immune response."

As implied, emotional support was operationalized as comments that extended empathy and compassion. This could often be found when there were accounts of personal experiences. Bible verses were shared as a means of offering hope. Commenters also thanked creators for sharing their story and offered prayers ([Table 8](#)). There was support from those living with other long-term illnesses, notably those with chronic fatigue syndrome or myalgic encephalomyelitis. Such discourse often led to community building in the comment section. This was particularly prevalent in the comment section of long hauler videos. To continue discussion, commenters asked follow-up questions regarding the progression of symptoms ([Table 9](#)). In addition, they sought other avenues to connect with and support each other ([Table 8](#)).

Skepticism, Misinformation, and Negatively Charged Comments on News and Medical Source Videos

We also observed a high frequency of negatively charged content, particularly in the comments for news and medical source videos. Skepticism regularly appeared in news-related content. Theories suggested by prominent politicians abounded, such as ivermectin as a cure for COVID-19. Many also criticized the credibility of the news sources and their supposed neutrality. News stations and reporters were, at times, labeled as people pushing liberal agendas and fear-mongering propaganda.

Misinformation and disinformation were major themes in both medical source and news source videos and comments. Some commenters felt that physicians were not sharing correct

information or were misinterpreting the information that they had received (Table 7). This was despite the fact that, in many medical source videos, there was ample time spent expounding on the credentials of guest speakers, perhaps in an attempt to boost credibility before information was shared. In contrast, some commenters shared the opposite—they appreciated the scientific approach taken by physicians as opposed to news sensationalism (Table 6).

Other negatively charged comments dealt with the lack of needs: lack of information, lack of visibility, and lack of improvement. In general, commenters sought more information from health care professionals (Table 7). On a related note, commenters expressed wishes for more visibility regarding PCC. Commenters noted that their symptoms did not improve even once given the vaccine.

On the basis of our sentiment analysis, news source videos received by far the greatest proportion of negative comments. When assessing the topics and themes that came up in comments under news source videos, criticism and sharing of misinformation were dominant. Many of the ideas shared by commenters reflected those of politicians. In these views, blame for the spread of COVID-19 and COVID-19–based restrictions was shifted onto China and liberal politicians. Vaccine hesitancy and opposition expressed by commenters were reiterated by some politicians as well. Some commenters appeared to experience extreme fear with regard to the vaccine. They mentioned that those administering the vaccine and treating long haulers had motives to kill (Table 5). This seems to shed light on the idea that, although many previously debunked sentiments of politicians were being repeated, there was a genuine fear of vaccines, the health care system, and some members of the government. The sentiment analysis of videos from medical sources revealed that only a smaller portion (528/1258, 41.97%) of the comments were negative.

Implications

The results of this study could help public health agencies, policy makers, organizations, and health researchers understand symptomatology and experiences related to PCC. The information includes a description of the diverse range of symptoms and informational and emotional needs of patients with PCC. This information can help public health professionals develop and implement effective interventions to manage PCC. Voices of Long Covid [39] is one campaign promoted by the US Department of Health and Human Services that emerged in November 2021 as a community for those with the syndrome. In addition to providing a forum for patients with PCC to share their experiences, the campaign offers resources for vaccinations and updates on developing research. The findings of this study demonstrate the potential of computational analysis of social media to provide insights and communication strategies regarding the public's responses to future health crises. This can be used to provide additional perspective and information to such campaigns.

As referenced in the NBC News article [35], there are patients with PCC who have been met with resistance by some medical professionals. For example, one patient felt that she was dismissed after explaining her PCC symptoms. This dilemma

has led to the creation of long hauler support groups on various social media platforms [35]. By mining YouTube, a rich source of our daily experiences, we began to uncover multifaceted challenges faced by long haulers. Our findings align with the experiences of patients who have lost work due to PCC and are unable to receive insurance coverage.

Limitations and Future Work

There are some limitations to this work. Our study was conducted on YouTube transcripts. In many cases, transcripts for YouTube videos are automatically generated. This means that the captioning process is imperfect and, at times, incorrect words were recorded instead of the words that the speakers said.

In addition, we only reviewed top-level comments related to our videos, and our analyses on comments does not reflect the full scope of the discourse in the comment section. Thus, we may be missing important insights from responses to the comments. Future studies should extend this study to include reactions to comments as well. Another limitation is that we cannot assume that the comments presented underneath the videos in our study are representative of all viewers. Many viewers do not comment on videos [40]; thus, their opinions are not captured.

It is difficult to detect sarcasm and linguistic nuances using LDA and sentiment analysis. Despite this, sarcasm is often used in everyday speech. Because of this, the computational models may have interpreted some texts differently from how they were originally intended.

Future research could focus on the longitudinal experience of long haulers to examine how they are perceived and their overall experience over time. Long hauler sentiments toward the health care system and physicians could potentially have changed over time. In addition, as more information has surfaced and more COVID-19 infections have likely led to more PCC cases, there may have been a change in the level of skepticism and distrust when it comes to long hauler experience. Longitudinal studies would be able to explore this shift in their experience. Future research could explore the effectiveness of various public health strategies in mitigating the impact of PCC considering potential changes in public awareness and understanding fostered by increased media coverage, including YouTube.

Regarding recent PCC treatments, we started our research before drugs such as Paxlovid received full Food and Drug Administration approval on November 2023 [41]. We collected the videos on November 1, 2021, which included videos made in August 2020 after the spread of COVID-19 and until October 2021. A future study should investigate how the availability of PCC treatments changed the perceptions, management, and psychological impact of PCC.

It is important to acknowledge that the commenters and video creators in our YouTube study may be subject to selection bias and have excluded certain geographic and demographic perspectives. These perspectives hold some weight in how public sentiment should be perceived [42-45]. However, >95% of the internet population spanning 88 countries regularly interacts with YouTube [46]. This highlights the potential opportunity for broader exploration.

Conclusions

In this study, we used topic modeling to investigate videos concerning PCC on YouTube. In addition, we assessed public responses to these videos by analyzing the comment section using sentiment analysis and topic modeling. We found that videos mostly focused on symptomatology, potential treatments, and sharing experiences. There was a range of response types, with news source videos receiving the highest proportion of negative comments and medical source videos receiving the lowest proportion of negative comments. Some were negative and often referenced conspiracy theories and distrust of the shared content. They also included negative experiences

regarding PCC symptoms and treatment. Positive comments were those that exhibited community building, sharing of information, and offering of support. This information, which is based on social media analyses, can assist public health professionals in comprehending the responses to PCC, includes a description of the diverse range of symptoms and informational and emotional needs of patients with PCC, and can help public health professionals develop and implement effective interventions to manage PCC. The findings of this study demonstrate the potential of computational analysis of social media to provide insights and communication strategies regarding the public's responses to future health crises.

Data Availability

The data sets generated during and analyzed during this study are available from the corresponding author on reasonable request.

Conflicts of Interest

None declared.

References

1. "Post-acute COVID-19 syndrome": COVID "long-haulers" suffering symptoms months after initial diagnosis. YouTube. URL: <https://www.youtube.com/watch?v=0gLmMPOHDwM> [accessed 2021-04-21]
2. Cutler DM. The costs of long COVID. JAMA Health Forum 2022 May 06;3(5):e221809 [FREE Full text] [doi: [10.1001/jamahealthforum.2022.1809](https://doi.org/10.1001/jamahealthforum.2022.1809)] [Medline: [36219031](https://pubmed.ncbi.nlm.nih.gov/36219031/)]
3. Rubin R. As their numbers grow, COVID-19 "long haulers" stump experts. JAMA 2020 Oct 13;324(14):1381-1383. [doi: [10.1001/jama.2020.17709](https://doi.org/10.1001/jama.2020.17709)] [Medline: [32965460](https://pubmed.ncbi.nlm.nih.gov/32965460/)]
4. Wisk LE, Nichol G, Elmore JG. Toward unbiased evaluation of postacute sequelae of SARS-CoV-2 infection: challenges and solutions for the long haul ahead. Ann Intern Med 2022 May;175(5):740-743. [doi: [10.7326/m21-4664](https://doi.org/10.7326/m21-4664)]
5. Pires L, Reis C, Mesquita Facão AR, Moniri A, Marreiros A, Drummond M, et al. Fatigue and mental illness symptoms in long COVID: protocol for a prospective cohort multicenter observational study. JMIR Res Protoc 2024 Jan 19;13:e51820 [FREE Full text] [doi: [10.2196/51820](https://doi.org/10.2196/51820)] [Medline: [38241071](https://pubmed.ncbi.nlm.nih.gov/38241071/)]
6. Lovelace BJ. Long Covid patients face medical debt after insurance denies claims. NBC News. 2023 Mar 9. URL: <https://www.nbcnews.com/health/health-news/long-covid-symptoms-treatment-insurance-coverage-rcna72012> [accessed 2024-05-14]
7. Jacques ET, Basch CH, Park E, Kollia B, Barry E. Long haul COVID-19 videos on YouTube: implications for health communication. J Community Health 2022 Aug 12;47(4):610-615 [FREE Full text] [doi: [10.1007/s10900-022-01086-4](https://doi.org/10.1007/s10900-022-01086-4)] [Medline: [35412189](https://pubmed.ncbi.nlm.nih.gov/35412189/)]
8. Oyebo O, Ndulue C, Adib A, Mulchandani D, Suruliraj B, Orji FA, et al. Health, psychosocial, and social issues emanating from the COVID-19 pandemic based on social media comments: text mining and thematic analysis approach. JMIR Med Inform 2021 Apr 06;9(4):e22734 [FREE Full text] [doi: [10.2196/22734](https://doi.org/10.2196/22734)] [Medline: [33684052](https://pubmed.ncbi.nlm.nih.gov/33684052/)]
9. Melton CA, Olusanya OA, Ammar N, Shaban-Nejad A. Public sentiment analysis and topic modeling regarding COVID-19 vaccines on the Reddit social media platform: a call to action for strengthening vaccine confidence. J Infect Public Health 2021 Oct;14(10):1505-1512 [FREE Full text] [doi: [10.1016/j.jiph.2021.08.010](https://doi.org/10.1016/j.jiph.2021.08.010)] [Medline: [34426095](https://pubmed.ncbi.nlm.nih.gov/34426095/)]
10. Mutanga MB, Abayomi A. Tweeting on COVID-19 pandemic in South Africa: LDA-based topic modelling approach. Afr J Sci Technol Innov Dev 2020 Oct 08;14(1):163-172. [doi: [10.1080/20421338.2020.1817262](https://doi.org/10.1080/20421338.2020.1817262)]
11. Li C, Jordan A, Song J, Ge Y, Park A. A novel approach to characterize state-level food environment and predict obesity rate using social media data: correlational study. J Med Internet Res 2022 Dec 13;24(12):e39340 [FREE Full text] [doi: [10.2196/39340](https://doi.org/10.2196/39340)] [Medline: [36512396](https://pubmed.ncbi.nlm.nih.gov/36512396/)]
12. Oduru T, Jordan A, Park A. Healthy vs. unhealthy food images: image classification of Twitter images. Int J Environ Res Public Health 2022 Jan 14;19(2):923 [FREE Full text] [doi: [10.3390/ijerph19020923](https://doi.org/10.3390/ijerph19020923)] [Medline: [35055742](https://pubmed.ncbi.nlm.nih.gov/35055742/)]
13. Jelodar H, Wang Y, Rabbani M, Ahmadi SB, Boukela L, Zhao R, et al. A NLP framework based on meaningful latent-topic detection and sentiment analysis via fuzzy lattice reasoning on YouTube comments. Multimed Tools Appl 2020 Sep 28;80(3):4155-4181. [doi: [10.1007/s11042-020-09755-z](https://doi.org/10.1007/s11042-020-09755-z)]
14. Leviner S. Recognizing the clinical sequelae of COVID-19 in adults: COVID-19 long-haulers. J Nurse Pract 2021 Sep;17(8):946-949 [FREE Full text] [doi: [10.1016/j.nurpra.2021.05.003](https://doi.org/10.1016/j.nurpra.2021.05.003)] [Medline: [33976591](https://pubmed.ncbi.nlm.nih.gov/33976591/)]

15. Thompson CM, Rhidenour KB, Blackburn KG, Barrett AK, Babu S. Using crowdsourced medicine to manage uncertainty on Reddit: the case of COVID-19 long-haulers. *Patient Educ Couns* 2022 Feb;105(2):322-330 [FREE Full text] [doi: [10.1016/j.pec.2021.07.011](https://doi.org/10.1016/j.pec.2021.07.011)] [Medline: [34281723](https://pubmed.ncbi.nlm.nih.gov/34281723/)]
16. Siegelman JN. Reflections of a COVID-19 long hauler. *JAMA* 2020 Nov 24;324(20):2031-2032. [doi: [10.1001/jama.2020.22130](https://doi.org/10.1001/jama.2020.22130)] [Medline: [33175108](https://pubmed.ncbi.nlm.nih.gov/33175108/)]
17. Basch CH, Park E, Kollia B, Quinones N. Online news coverage of COVID-19 long haul symptoms. *J Community Health* 2022 Apr 03;47(2):306-310 [FREE Full text] [doi: [10.1007/s10900-021-01053-5](https://doi.org/10.1007/s10900-021-01053-5)] [Medline: [34860328](https://pubmed.ncbi.nlm.nih.gov/34860328/)]
18. Uddin SM, Albert A, Tamanna M, Alsharif A. YouTube as a source of information: early coverage of the COVID-19 pandemic in the context of the construction industry. *Construct Manage Econ* 2023 Jan 02;41(5):402-427. [doi: [10.1080/01446193.2022.2162096](https://doi.org/10.1080/01446193.2022.2162096)]
19. Leading countries based on YouTube audience size as of January 2024. Statista. URL: <https://www.statista.com/statistics/280685/number-of-monthly-unique-youtube-users/> [accessed 2022-07-18]
20. Syed-Abdul S, Fernandez-Luque L, Jian WS, Li YC, Crain S, Hsu MH, et al. Misleading health-related information promoted through video-based social media: anorexia on YouTube. *J Med Internet Res* 2013 Feb 13;15(2):e30 [FREE Full text] [doi: [10.2196/jmir.2237](https://doi.org/10.2196/jmir.2237)] [Medline: [23406655](https://pubmed.ncbi.nlm.nih.gov/23406655/)]
21. McLellan A, Schmidt-Waselenchuk K, Duerksen K, Woodin E. Talking back to mental health stigma: an exploration of YouTube comments on anti-stigma videos. *Comput Hum Behav* 2022 Jun;131:107214. [doi: [10.1016/j.chb.2022.107214](https://doi.org/10.1016/j.chb.2022.107214)]
22. Aslam AB, Syed ZS, Khan MF, Baloch A, Syed MS. Leveraging natural language processing for public health screening YouTube: a COVID-19 case study. arXiv Preprint posted online June 1, 2023 [FREE Full text]
23. Serrano JC, Papakyriakopoulos O, Hegelich S. NLP-based feature extraction for the detection of COVID-19 misinformation videos on YouTube. In: Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020. 2020 Presented at: ACL 2020; July 9-10, 2020; Online.
24. Choi B, Kim H, Huh-Yoo J. Seeking mental health support among college students in video-based social media: content and statistical analysis of YouTube videos. *JMIR Form Res* 2021 Nov 11;5(11):e31944 [FREE Full text] [doi: [10.2196/31944](https://doi.org/10.2196/31944)] [Medline: [34762060](https://pubmed.ncbi.nlm.nih.gov/34762060/)]
25. Yan X, Guo J, Lan Y, Cheng X. A bitern topic model for short texts. In: Proceedings of the 22nd international conference on World Wide Web. 2013 Presented at: WWW '13; May 13-17, 2013; Rio de Janeiro, Brazil. [doi: [10.1145/2488388.2488514](https://doi.org/10.1145/2488388.2488514)]
26. Schwarz C. Ldagibbs: a command for topic modeling in stata using latent dirichlet allocation. *Stata J* 2018 Mar 01;18(1):101-117. [doi: [10.1177/1536867x1801800107](https://doi.org/10.1177/1536867x1801800107)]
27. Mimno D, Wallach HM, Talley E, Leenders M, McCallum A. Optimizing semantic coherence in topic models. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. 2011 Presented at: EMNLP '11; July 27-31, 2011; Edinburgh, UK. [doi: [10.3115/1699571.1699627](https://doi.org/10.3115/1699571.1699627)]
28. Corbin J, Strauss A. Basics of Qualitative Research: Techniques and Procedures for Developing Grounded Theory. Thousand Oaks, CA: SAGE Publications; 1990.
29. Kwon S, Park A. Examining thematic and emotional differences across Twitter, Reddit, and YouTube: the case of COVID-19 vaccine side effects. *Comput Human Behav* 2023 Jul;144:107734 [FREE Full text] [doi: [10.1016/j.chb.2023.107734](https://doi.org/10.1016/j.chb.2023.107734)] [Medline: [36942128](https://pubmed.ncbi.nlm.nih.gov/36942128/)]
30. Hutto C, Gilbert E. VADER: a parsimonious rule-based model for sentiment analysis of social media text. *Proc Int AAAI Conf Web Soc Med* 2014;8(1):216-225. [doi: [10.1609/icwsm.v8i1.14550](https://doi.org/10.1609/icwsm.v8i1.14550)]
31. Ahmed A, Aziz S, Khalifa M, Shah U, Hassan A, Abd-Alrazaq A, et al. Thematic analysis on user reviews for depression and anxiety chatbot apps: machine learning approach. *JMIR Form Res* 2022 Mar 11;6(3):e27654 [FREE Full text] [doi: [10.2196/27654](https://doi.org/10.2196/27654)] [Medline: [35275069](https://pubmed.ncbi.nlm.nih.gov/35275069/)]
32. Immunophenotyping. Stanford Health Care. URL: <https://stanfordhealthcare.org/medical-tests/i/immunophenotyping.html> [accessed 2024-05-14]
33. Trihandini I, Muhta M, Sakti DA, Erlianti CP. Effects of long-haul COVID on the health-related quality of life among recovered hospitalized patients. Research Square Preprint posted online March 28, 2022 [FREE Full text] [doi: [10.21203/rs.3.rs-1478232/v1](https://doi.org/10.21203/rs.3.rs-1478232/v1)]
34. Park A, Conway M. Harnessing Reddit to understand the written-communication challenges experienced by individuals with mental health disorders: analysis of texts from mental health communities. *J Med Internet Res* 2018 Apr 10;20(4):e121 [FREE Full text] [doi: [10.2196/jmir.8219](https://doi.org/10.2196/jmir.8219)] [Medline: [29636316](https://pubmed.ncbi.nlm.nih.gov/29636316/)]
35. Bellamy C, Adams C. Black Covid long-haulers felt invisible to the health care system, so they formed their own support groups. NBC News. 2022 Aug 28. URL: <https://www.nbcnews.com/news/nbcblk/black-covid-long-haulers-felt-invisible-health-care-system-formed-supp-rcna44468> [accessed 2024-05-14]
36. Pfaff ER, Madlock-Brown C, Baratta JM, Bhatia A, Davis H, Girvin A, et al. Coding long COVID: characterizing a new disease through an ICD-10 lens. *BMC Med* 2023 Mar 16;21(1):58 [FREE Full text] [doi: [10.1186/s12916-023-02737-6](https://doi.org/10.1186/s12916-023-02737-6)] [Medline: [36793086](https://pubmed.ncbi.nlm.nih.gov/36793086/)]

37. Khullar D, Zhang Y, Zang C, Xu Z, Wang F, Weiner MG, et al. Racial/ethnic disparities in post-acute sequelae of SARS-CoV-2 infection in New York: an EHR-based cohort study from the RECOVER program. *J Gen Intern Med* 2023 Apr 16;38(5):1127-1136 [FREE Full text] [doi: [10.1007/s11606-022-07997-1](https://doi.org/10.1007/s11606-022-07997-1)] [Medline: [36795327](https://pubmed.ncbi.nlm.nih.gov/36795327/)]
38. Health literacy: report of the Council on Scientific Affairs. Ad Hoc Committee on Health Literacy for the Council on Scientific Affairs, American Medical Association. *JAMA* 1999 Mar 10;281(6):552-557. [Medline: [10022112](https://pubmed.ncbi.nlm.nih.gov/10022112/)]
39. Voices of long Covid. Resolve to Save Lives. URL: <https://voicesoflongcovid.org/> [accessed 2024-03-08]
40. Soukup PA. Looking at, through, and with YouTube. *Commun Res Trends* 2014;33(3):3-34 [FREE Full text]
41. Pfizer's PAXLOVID™ receives FDA approval for adult patients at high risk of progression to severe COVID-19. Pfizer. 2023 May 25. URL: <https://www.pfizer.com/news/press-release/press-release-detail/pfizers-paxlovidtm-receives-fda-approval-adult-patients> [accessed 2024-03-08]
42. Padilla JJ, Kavak H, Lynch CJ, Gore RJ, Diallo SY. Temporal and spatiotemporal investigation of tourist attraction visit sentiment on Twitter. *PLoS One* 2018 Jun 14;13(6):e0198857 [FREE Full text] [doi: [10.1371/journal.pone.0198857](https://doi.org/10.1371/journal.pone.0198857)] [Medline: [29902270](https://pubmed.ncbi.nlm.nih.gov/29902270/)]
43. Mitchell L, Frank MR, Harris KD, Dodds PS, Danforth CM. The geography of happiness: connecting Twitter sentiment and expression, demographics, and objective characteristics of place. *PLoS One* 2013 May 29;8(5):e64417 [FREE Full text] [doi: [10.1371/journal.pone.0064417](https://doi.org/10.1371/journal.pone.0064417)] [Medline: [23734200](https://pubmed.ncbi.nlm.nih.gov/23734200/)]
44. Hussein E, Juneja P, Mitra T. Measuring misinformation in video search platforms: an audit study on YouTube. *Proc ACM Hum Comput Interact* 2020 May 29;4(CSCW1):1-27. [doi: [10.1145/3392854](https://doi.org/10.1145/3392854)]
45. Gore RJ, Diallo S, Padilla J. You are what you tweet: connecting the geographic variation in America's obesity rate to Twitter content. *PLoS One* 2015;10(9):e0133505 [FREE Full text] [doi: [10.1371/journal.pone.0133505](https://doi.org/10.1371/journal.pone.0133505)] [Medline: [26332588](https://pubmed.ncbi.nlm.nih.gov/26332588/)]
46. Osman W, Mohamed F, Elhassan M, Shoufan A. Is YouTube a reliable source of health-related information? A systematic review. *BMC Med Educ* 2022 May 19;22(1):382 [FREE Full text] [doi: [10.1186/s12909-022-03446-z](https://doi.org/10.1186/s12909-022-03446-z)] [Medline: [35590410](https://pubmed.ncbi.nlm.nih.gov/35590410/)]

Abbreviations

LDA: latent Dirichlet allocation

PCC: post-COVID-19 condition

RQ: research question

VADER: Valence Aware Dictionary and Sentiment Reasoner

Edited by H Liu; submitted 14.11.23; peer-reviewed by R Gore, A Wahbeh; comments to author 14.02.24; revised version received 02.04.24; accepted 06.04.24; published 03.06.24.

Please cite as:

Jordan A, Park A

Understanding the Long Haulers of COVID-19: Mixed Methods Analysis of YouTube Content

JMIR AI 2024;3:e54501

URL: <https://ai.jmir.org/2024/1/e54501>

doi: [10.2196/54501](https://doi.org/10.2196/54501)

PMID: [38875666](https://pubmed.ncbi.nlm.nih.gov/38875666/)

©Alexis Jordan, Albert Park. Originally published in JMIR AI (<https://ai.jmir.org>), 03.06.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR AI, is properly cited. The complete bibliographic information, a link to the original publication on <https://www.ai.jmir.org/>, as well as this copyright and license information must be included.

Original Paper

Role of Ethics in Developing AI-Based Applications in Medicine: Insights From Expert Interviews and Discussion of Implications

Lukas Weidener¹, BSc, Dr med; Michael Fischer¹, PhD

Research Unit for Quality and Ethics in Health Care, UMIT TIROL – Private University for Health Sciences and Health Technology, Hall in Tirol, Austria

Corresponding Author:

Lukas Weidener, BSc, Dr med

Research Unit for Quality and Ethics in Health Care

UMIT TIROL – Private University for Health Sciences and Health Technology

Eduard-Wallnöfer-Zentrum 1

Hall in Tirol, 6060

Austria

Phone: 43 17670491594

Email: lukas.weidener@edu.umit-tirol.at

Abstract

Background: The integration of artificial intelligence (AI)-based applications in the medical field has increased significantly, offering potential improvements in patient care and diagnostics. However, alongside these advancements, there is growing concern about ethical considerations, such as bias, informed consent, and trust in the development of these technologies.

Objective: This study aims to assess the role of ethics in the development of AI-based applications in medicine. Furthermore, this study focuses on the potential consequences of neglecting ethical considerations in AI development, particularly their impact on patients and physicians.

Methods: Qualitative content analysis was used to analyze the responses from expert interviews. Experts were selected based on their involvement in the research or practical development of AI-based applications in medicine for at least 5 years, leading to the inclusion of 7 experts in the study.

Results: The analysis revealed 3 main categories and 7 subcategories reflecting a wide range of views on the role of ethics in AI development. This variance underscores the subjectivity and complexity of integrating ethics into the development of AI in medicine. Although some experts view ethics as fundamental, others prioritize performance and efficiency, with some perceiving ethics as potential obstacles to technological progress. This dichotomy of perspectives clearly emphasizes the subjectivity and complexity surrounding the role of ethics in AI development, reflecting the inherent multifaceted nature of this issue.

Conclusions: Despite the methodological limitations impacting the generalizability of the results, this study underscores the critical importance of consistent and integrated ethical considerations in AI development for medical applications. It advocates further research into effective strategies for ethical AI development, emphasizing the need for transparent and responsible practices, consideration of diverse data sources, physician training, and the establishment of comprehensive ethical and legal frameworks.

(JMIR AI 2024;3:e51204) doi:[10.2196/51204](https://doi.org/10.2196/51204)

KEYWORDS

artificial intelligence; AI; medicine; ethics; expert interviews; AI development; AI ethics

Introduction

Background

Artificial intelligence (AI) has been considered a key technology in medical advancement for several years [1]. Recent developments in AI, exemplified by the broad availability and widespread use of advanced AI-based chat applications, such as ChatGPT, have underscored the capabilities of technology

[2]. This study specifically focuses on AI-based applications in medicine, highlighting the importance of ethics in their development, with an emphasis on the role of developers. Considering the inherent complexities associated with AI and its applications in medicine along with the multifaceted nature of AI ethics, this introduction aims to provide a comprehensive foundation for this publication.

Artificial Intelligence

Early definitions of AI, such as by McCarthy et al [3], primarily focused on the potential for machines to simulate all facets of human intelligence: "...the basis of the conjecture that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it." Newer definitions, such as the one from the European Parliament, expand this scope and describe AI as "the ability of a machine to display a range of humanlike capabilities, including reasoning, learning, planning, and creativity," encompassing a broader spectrum of intelligent behaviors [4].

Following the evolving definitions of AI, the term broadly encompasses various technologies, each with unique characteristics and applications. The scientific community commonly categorizes these technologies as "strong AI" and "weak AI" [5]. "Strong AI" refers to systems whose cognitive capabilities are comparable with human intelligence across a wide range of tasks and contexts [5]. However, most current applications, particularly in medicine, are categorized as "weak AI." This category includes systems designed to perform specific tasks using cognitive abilities comparable with those of humans but within a limited scope [6]. Within the category of "weak AI," 2 primary subfields are prominent: expert systems and machine learning (ML) [6]. Expert systems, categorized under "symbolic AI," operate based on predefined rules and instructions set by human experts [7]. In contrast, ML represents the "statistical AI" subfield [8]. ML focuses on pattern recognition within large data sets, enabling the system to learn and make predictions or decisions based on the data [9]. A notable example of such advancements in "statistical AI" is the development of large language models, such as ChatGPT, which demonstrate the evolving capabilities of AI in understanding and generating humanlike text, offering new possibilities, and raising unique ethical considerations in their application [10].

Despite the significant technological advances in the field of AI and, in particular, "weak AI," "strong AI," which would entail cognitive abilities on par with human intelligence across diverse areas, remains largely theoretical with no substantial application in medicine to date [11]. Therefore, "weak AI" will be the foundation of this publication, specifically focusing on the development and associated ethical considerations of "symbolic AI" and "statistical AI" applications in medicine.

AI in Medicine

The technological advancements and capabilities of AI in medicine, as exemplified by a range of AI-based applications such as ML algorithms and expert systems, are anticipated to transform various aspects of health care, such as diagnostics or personalized treatment planning [1].

For example, ML algorithms, a key subset of "statistical AI," are of particular interest in medicine because of their capability to analyze large data sets, including a wide array of medical images such as x-rays, magnetic resonance imaging, computed tomography, and dermatological photographs [8]. In radiology, ML algorithms enhance image interpretation by identifying the features associated with specific pathologies. For instance, in mammography, ML assists radiologists in detecting

microcalcifications and subtle changes in the breast tissue, which may indicate the early stages of breast cancer [12]. Similarly, in dermatology, ML-powered tools analyze photographic data of skin lesions and moles, thereby providing critical diagnostic insights [13]. By distinguishing between benign and malignant lesions with high accuracy, the early detection of skin cancer can be improved. The integration of ML in image-based diagnostics can not only enhance diagnostic accuracy but also have the potential to speed up the diagnostic process [8]. This reduction in analysis time leads to quicker diagnostic outcomes, enabling earlier intervention and treatment, which are crucial for improving patient care [14].

Expert systems in medicine, a subfield of "symbolic AI," are primarily exemplified by Clinical Decision Support Systems (CDSS) [15]. By leveraging predefined rules and knowledge from medical experts, these systems can provide recommendations for diagnosis and therapy options, potentially enhancing the decision-making process in clinical settings [16]. CDSS often use information from various sources, such as electronic health records, patient history, and latest medical research, to offer evidence-based suggestions. In addition to offering diagnostic and treatment guidance, CDSS can play a significant role in identifying potential adverse drug events, which is a critical aspect of patient safety [16]. By cross-referencing a patient's current medications with the proposed treatments, CDSS can alert health care providers to possible drug-drug interactions, allergic reactions, or contraindications based on the patient's medical history or known conditions [15].

In addition to diagnostic and decision support applications, AI contributes to other areas of medicine, such as medical research and drug development. In medical research, AI algorithms are used to analyze complex information, such as genetic, environmental, and lifestyle data, which can be used for personalized medical approaches, enabling more targeted therapies based on individual patient profiles [17]. Furthermore, AI can be used to identify potential therapeutic compounds more quickly and efficiently than traditional methods [18]. AI systems can simulate and predict how different compounds interact with biological targets, thereby reducing the time and cost of drug trials. This capability is particularly crucial in rapidly responding to emerging global health challenges, such as the development of vaccines and treatments for new diseases [18]. Furthermore, although AI-based chat applications, such as ChatGPT, have not been specifically developed for use in medicine, they possess extensive medical knowledge, making their potential application in various medical contexts a subject of increasing interest [2]. Although advancements in the field of AI can offer transformative benefits for medicine, they also introduce new ethical considerations and challenges that warrant attention [19,20].

AI Ethics

AI ethics can be defined as "a set of values, principles, and techniques that employ widely accepted standards of right and wrong to guide moral conduct in the development and use of AI technologies" [21]. Although this definition does not specifically focus on or include the field of medicine, it

emphasizes the importance of values and principles in the development of AI technologies. In medicine, the fundamental principles of medical ethics formulated by Beauchamp and Childress—autonomy, nonmaleficence, beneficence, and justice—are of paramount influence and relevance [22].

The principle of autonomy emphasizes respecting patients' rights to make informed decisions regarding their own health. In the context of AI-based applications in medicine, the principle of autonomy often refers to the development of technologies that support and enhance patient decision-making while maintaining transparency, explainability, and accountability [23,24]. This also refers to the development of AI-based applications that not only provide accurate diagnostic and treatment recommendations but also present their findings in a manner that is understandable and useful for both patients and health care professionals. The principle of nonmaleficence, emphasizing the commitment to do no harm, has become increasingly important in the context of growing role of AI in health care. Adhering to this principle requires the establishment of stringent safety protocols and comprehensive testing of AI technologies to prevent unintended consequences, such as biases in decision-making that could lead to misdiagnosis or unequal treatment of patients [24].

Bias in AI systems, particularly in medical applications, is a significant concern. For instance, ML algorithms used in image-based diagnostics, such as those used in radiology or dermatology, may develop biases based on the data they are trained on [25]. If these algorithms are primarily trained on data sets that lack diversity, they might be less accurate in diagnosing conditions in patient populations that are underrepresented in the training data [26]. This can lead to disparities in diagnostic accuracy and effectiveness, potentially harming certain groups of patients [16,27]. Similarly, in CDSS, which rely on predefined rules and medical knowledge, there is a risk of inherent biases being transferred into the system. If the input data or rules within these systems reflect historical biases or unequal treatment practices, the CDSS might perpetuate these issues, leading to recommendations that are not equitable or appropriate for all patients [16].

Addressing the challenges related to autonomy and nonmaleficence is fundamental for ensuring that AI in medicine aligns with the principles of beneficence and justice. The principle of beneficence, or acting in the best interests of the patient, emphasizes that AI-based applications in medicine should be developed with the primary goal of improving patient outcomes and enhancing quality of care [23]. Finally, the principle of justice requires that AI technologies in health care promote fairness and equity. This means ensuring equitable access to the benefits of AI advancements regardless of a patient's socioeconomic status or background [24].

In light of these ethical principles, the role of developers in creating AI-based applications in medicine has become critically important. Developers bear a particular responsibility to ensure that the design and implementation of these technologies adhere to the ethical standards outlined by autonomy, nonmaleficence, beneficence, and justice [28]. A deep understanding and awareness of the ethical implications during the development

process are essential, as the principles and guidelines frequently discussed in the current literature should be integrated from the early stages of AI application development [29,30]. This integration is not just theoretical but requires practical implementation and consistent consideration throughout the development process of AI-based applications in health care [31]. Despite the crucial role that developers play in embedding these ethical principles into AI technologies, there remains a gap in the literature regarding how developers perceive and prioritize ethics in their work [32,33]. Addressing this gap is essential for ensuring the responsible development and use of AI in medicine and aligning technological advancements with the core values of medical ethics.

Objective

The field of AI-based medical applications is rapidly advancing; however, a significant gap remains in understanding how ethical considerations are integrated into this development process. Recognizing the frequent calls in the literature for consistent inclusion of ethics in AI development, this study aimed to bridge this gap by exploring the perceptions, priorities, and conflicts related to ethics among AI experts. Specifically, this study sought to answer the following questions:

- How do AI experts perceive the role of ethics in the development of AI-based medical applications?
- How do AI experts perceive the relationship between ethical considerations and the technical development of AI-based applications in medicine?

The primary objective of this study is not only to answer these critical questions but also to provide an in-depth discussion of the results, particularly focusing on the associated ethical implications. This exploration is vital for understanding how ethical considerations can be more effectively integrated into the development of AI technologies in medical settings with the aim of contributing to the responsible and beneficial advancement of this field.

Methods

To address the study's objective, a secondary analysis of the exploratory expert interviews was performed using qualitative content analysis. These interviews were initially conducted to explore the essential knowledge and understanding of AI in medicine, with the aim of specifying teaching content on AI for medical education [34].

Ethical Considerations

Ethics approval was granted by the Research Committee for Scientific Ethical Questions of the UMIT TIROL—Private University for Health Sciences and Health Technology, Hall in Tirol, Austria, for both the initial data collection and secondary analysis of the data relevant to this study (approval number: 3181; January 16, 2023).

The methodology and reporting of the research findings in this study were guided by the Standards for Reporting Qualitative Research to ensure clarity and transparency [35].

Expert Characteristics

Of the 12 experts included in the primary research study, 7 met the inclusion criteria for this study and provided information relevant to the study objective. For this secondary analysis, individuals were defined as experts if they had been engaged in the research or practical development of AI-based applications in medicine for at least 5 years. In this regard, 4 experts were involved in the development of AI-based applications as part of their research activities (eg, researchers at the German Research Center for Artificial Intelligence, professor for medical informatics), such as enhanced AI-assisted imaging. The remaining 3 experts were primarily engaged in the practical development of various AI-based applications for use in medicine (eg, voice recognition in hospitals or assistance in diagnosis in medical practices) as part of their main professional

activities in the private sector (eg, software development). Additional inclusion criteria were sufficient language skills (German) and consent for the transcription of the interviews and their evaluation. All 7 participating experts were situated and working in Germany, providing a national perspective on the development of AI in medicine. Of the 7 experts included in this secondary analysis, 6 identified as male and 1 (E2) identified as female. Although all experts met the inclusion criteria of being engaged in research or the practical development of AI-based applications in medicine for at least 5 years, 3 experts (E1, E2, and E4) had more than 10 years of professional experience in the relevant field. In addition, 3 experts had more than 15 years of experience in the field of research and practical development of AI-based applications (E3, E5, and E7). [Table 1](#) presents a detailed overview of the experts' characteristics included in this study.

Table 1. Characteristics of the experts included in the secondary analysis.

Expert number	Professional position	Domain of expertise
E1	Research and development (AI ^a)	Machine learning in pathology
E2	Data scientist	AI in radiology
E3	Senior software developer	Clinical Decision Support Systems
E4	Research and development (AI)	AI in cancer diagnosis
E5	Professor for medical informatics	Natural language processing in medicine
E6	Data scientist	AI-assisted voice analysis for diagnosis
E7	Senior software developer	Clinical Decision Support Systems

^aAI: artificial intelligence.

Data Collection

In the initial data collection phase of the primary study, experts were recruited primarily via email. In addition, participants were asked to recommend other potential experts for the interviews, thereby expanding the recruitment network. This direct recommendation approach enabled the inclusion of 2 additional experts in the primary study. Before the interviews were recorded, the experts were informed about the study and the associated data protection regulations during recruitment and at the beginning of the interviews. All interviews were conducted using a video service provider (Cisco Webex Meetings) and were recorded on an audio basis (manual recording via an analog dictation device; average interview length 34.02, SD 4.1 minutes).

To ensure the protection of all collected and generated data, they were stored offline on a password-protected storage device in a lockable cabinet, with access limited to the researcher. The anonymized data will be stored for 10 years following the date of collection to enable reproducibility and deleted after to ensure confidentiality. All participating experts explicitly consented to both the initial analysis and the use of their data for future research purposes, as in the case of this study.

Data Analysis

The expert interviews were transcribed using the transcription software f4transcript and anonymized according to the transcription rules of Dresing and Pehl [36]. The evaluation of

the collected data was conducted with software support (QCAmap, version 1.2.0; Microsoft Excel, version 16.66) and was rule based according to the methodology of qualitative content analysis by Mayring (inductive procedure) [37]. Relevant categories were defined directly from the material and were controlled or revised after viewing 40% of the material. After defining the categories, the entire material was reviewed, and relevant text passages were assigned to the respective main and subcategories.

The interviews were conducted and analyzed in German. For this publication, all identified and relevant text passages were translated into the English language. The primary research team conducted the initial translation, followed by a review and revision by a professional academic translator.

It is noteworthy that the data analysis in this study was guided by the research team's perspective and understanding of ethics. As such, the interpretation of the data and subsequent conclusions are shaped by the team's affiliation with the research unit for quality and ethics in health care. Consequently, ethical considerations, particularly in health care and medicine as well as in the development and application of AI technologies in these fields, are considered important. The emphasis on ethics should be considered when interpreting the results of this study.

Furthermore, the aspect of theoretical saturation in this secondary analysis warrants detailed discussion. Given its distinct objectives, this study selectively used interviews with 7 of the 12 experts, chosen based on the specific inclusion

criteria of engagement in research or practical development of AI-based applications in medicine for over 5 years. The remaining 5 experts from the primary study, who primarily focused on teaching and research without a direct emphasis on developing AI-based applications for medicine, did not meet the inclusion criteria for this secondary analysis. This selection, inherent to the secondary nature of the data, led to a focused but relatively limited breadth in certain areas, resulting in incomplete saturation in the 2 subcategories. Specifically, the subcategories of “Data Protection” (section *Subcategory 3: Data Protection*) and “Demands” (section *Subcategory 3: Data Protection*) demonstrated incomplete saturation, each substantiated by only a single reference. In contrast, theoretical saturation for the other categories can be assumed, given the multiple references that support the established themes and the lack of new insights, suggesting the need for additional categories.

Acknowledging this limitation is crucial, particularly in the context of future research opportunities aimed at more comprehensively exploring these underrepresented areas. However, the reliability of the results extends beyond the theoretical saturation. It is also underscored by the expertise and extensive experience of the participating experts, each with at least 5 years of AI research or practical development in medicine. Their profound insights, combined with the systematic and iterative analysis methodology, ensured that the extracted themes were representative and comprehensive, despite the gaps noted in certain subcategories. Consequently, although the findings in the “Data Protection” and “Demands” categories might benefit from further exploration in future studies, the current analysis offers a robust and insightful understanding of the primary themes related to ethical considerations in AI development for medical applications.

Textbox 1. Overview of the 3 main categories with a total of 7 subcategories from the analysis of interviews with experts in artificial intelligence.

<p>Essential foundation</p> <ul style="list-style-type: none"> • Awareness • Consequences • Data protection <p>Results in the foreground</p> <ul style="list-style-type: none"> • Performance • Economic efficiency <p>Obstacle to progress</p> <ul style="list-style-type: none"> • Demands • Blockade

First Main Category: Essential Foundation

As part of the first main category (“essential foundation”), all the statements defining ethics as an essential basis for the development of AI-based applications in medicine were summarized.

To ensure detailed and comprehensive data collection, a semistructured interview guideline was used for primary data collection. This interview guideline included questions directly related to the study’s objectives and incorporated both immanent and exmanent questioning. Reflecting the research team’s focus on ethics in health care and medicine, the semistructured interview guidelines incorporated 2 questions directly relevant to the study’s objectives:

- How do you perceive the role of ethics in the context of AI-based medical applications?
- What are your experiences with ethical considerations and the development of AI-based applications in medicine?

In addition to the 2 questions directly addressing the objective of this study, an interview guideline was constructed to promote openness by emphasizing the immanent and exmanent questions. Examples of the questions used are as follows:

- You have mentioned the challenge of integrating ethics into AI development. Could you elaborate on the specific ethical considerations you find most relevant in this context?
- In your view, who should bear responsibility for the ethical issues in AI-based applications—users or developers?

Using both direct and immanent as well as exmanent question types, the interviews aimed to provide an in-depth exploration of the topic of AI in medicine, including the development of AI-based applications for use in medicine.

Results

Overview

On the basis of the qualitative content analysis of the expert interviews, 3 main categories with 7 subcategories were defined using anchor examples. [Textbox 1](#) provides an overview of the main categories and subcategories defined.

Subcategory 1: Awareness

The first subcategory, “awareness,” highlights the relevance of ethics in development because of the potential dangers and consequences associated with AI:

Because AI is a sharp weapon, [unintelligible] it can be sharpened arbitrarily. But it must be used wisely. And I think one of the biggest difficulties is to anticipate, what does it actually mean when we develop this? [...] this anticipatory ethical question is extremely difficult. [E1; quote A.1]

This subcategory emphasizes the importance of developers being cognizant of the potential uses and challenges that may arise with the subsequent implementation of AI-based applications in medical settings. An additional perspective further reinforces this view:

If we develop something, we always think the application will be used as anticipated in the clinical setting. But we can never be sure, and developers need to be aware of this. [E5; quote A.2]

Subcategory 2: Consequences

The second subcategory “consequences,” further emphasizes the importance of ethics in practical development and an associated awareness to prevent consequences such as biases in the data or other potential forms of discrimination from being incorporated into the application:

I think everyone working with AI, especially the field of medicine or [unintelligible], should think of potential consequences involved with it. This does not only include the development teams or companies, but rather anyone. [E4; quote A.3]

Although the previous quote offers a broad view of the ethical considerations in AI for medicine, the next quote from a different expert highlights specific concerns, such as bias and its potential harm to patients:

Yes, well, ethics is super important. [...] Well, when we talk about this bias, when we talk about these false negatives, it's very important. [...] I am mostly afraid bias. Bias could really harm patients with potentially fatal outcomes. To limit the risk of any bias, we have ongoing discussions in the team. [E5; quote A.4]

Subcategory 3: Data Protection

The importance of ethics is also highlighted in terms of the general use of human data in the development of AI-based applications, thereby forming the foundation of the third subcategory:

Well, we actually have this discussion all the time. We at [...] have an ethics working group, for ethical processing and also [unintelligible] and equality. These aspects are always there, especially when you are working with data and people, [unintelligible] data generated by people. [E4; quote A.5]

Second Main Category: Results in the Foreground

In the context of the second main category, all statements from the experts are summarized, in which the “Results are in the foreground” of the development of AI-based algorithms.

Subcategory 1: Performance

The following quote from the analysis of the third expert interview reflects the result-oriented nature of the development of AI-based applications in medicine, which underlies the formation of the first subcategory:

For me insofar, and I also indirectly deal with it [ethics], but for me it does not represent the first thing. So, if it's for me, let us say, I want to set up a system first, then it's also about, I want to set up the system. Ethical aspects do not play a role for me. [...] sounds mean now, but when an IT specialist first trains his models, it's just about, as banal as it sounds, it's just about achieving good performance first. [E3; quote B.1]

This result and performance-driven perspective was echoed by another expert, who highlighted the competitive nature of AI development:

But I also believe that there are, let me say, more important things than ethics. Especially with the increased interest in AI, the competition is hard. [...] Developers as well as the applications do need to perform well. [E2; quote B.2]

These statements collectively underscore a tendency within the industry to prioritize performance metrics, which may occasionally overshadow ethical considerations in the drive to advance and remain competitive in the rapidly evolving AI sector.

Subcategory 2: Economic Efficiency

The subordinate significance of ethics in performance is also clarified by the following statement in the second subcategory:

I think companies that are in competition, even if they don't mean it badly, still have the market economic pressure to deliver results, and this can certainly also lead to losing sight of maintaining some ethical boundaries that one would better keep a careful eye on. [E6; quote B.3]

This sentiment is reiterated by another expert who highlights the financial imperatives driving company behavior:

In the end, earning money and making a profit is important to anyone being paid by companies. [...] This might be different in academia, like research, but we all need to focus on creating a product that does financially well, and not trying to be ethically correct. [Interview E2; quote B.4]

These perspectives elucidate the conflict that experts perceive between economic efficiency and ethical conduct in the development of AI-based medical applications.

Third Main Category: Obstacle to Progress

The third main category summarizes statements from experts who view ethics as an “obstacle to (technological) progress.”

Subcategory 1: Demands

As part of the first subcategory, the “Demands” of ethics are viewed as potential barriers that can stand in the way of AI

technology and the technological progress of AI-based applications in medicine:

I always find it a bit difficult to draw this line between these ethical demands and the limits that then really stand in the way of technology and progress. [E6; quote C.1]

Subcategory 2: Blockade

The perception that ethics can not only hinder current development but also impede future progress in AI forms the basis of the “Blockade” subcategory. This is exemplified by the following statement:

Please stop bothering me on the topic of ethics in AI. It blocks at all corners and edges. [...] Yes, but if I don't start, how should someone else continue in ten, 20 years so that something comes out of it? [E7; quote C.2]

The aforementioned quote illustrates a dismissive attitude toward ethics as part of the development process of AI-based applications in medicine and thus clarifies the assessment of ethics as an obstacle to (technological) progress. This perspective was reinforced by an additional quote from another expert:

I have no doubt that ethics is important, but it does not help the technological progress of AI. [...] Ethics can really prevent any meaningful advancement. [E6; quote C.3]

Together, these quotes highlight a critical perspective within the AI development community, where ethical concerns, although important, are sometimes seen as obstructions to both immediate technological development and long-term innovation in AI.

Discussion

Principal Findings

The results of the qualitative content analysis revealed a nuanced spectrum of expert opinions regarding the role of ethics in AI development for medical applications. Initially, in the “essential foundation” category, a consensus was observed among experts (eg, E1 and E5) on the foundational importance of ethics in AI development. This consensus on the foundational role of ethics is based on an understanding of AI's potential risks and consequences of AI, as exemplified by the anticipatory ethical questions posed by E1 (quote A.1) and the emphasis on uncertainty in application outcomes noted by E5 (quote A.2).

Within the “results in the foreground” category, a shift in perspective becomes apparent. Experts, such as E3 and E2, express views that prioritize performance and competitive outcomes over ethical considerations (quotes B.1 and B.2). This shift suggests a conflict between ethical integrity and market-driven objectives, with the latter often taking precedence in the fast-paced competitive landscape of AI development.

In the “obstacle to progress” category, the tension between ethical demands and technological advancement is further articulated. Expert E6, for instance, acknowledged the difficulty

of reconciling ethical demands with the limits imposed on technology and progress (quote C.1). This sentiment is echoed by expert E7, who expresses frustration with ethics perceived as a blockade of development (quote C.2). These perspectives underscore a critical view within the AI development community, where ethical concerns, although recognized as important, are sometimes seen as obstacles to immediate technological development and long-term innovation.

This variety of opinions, ranging from viewing ethics as foundational to considering them as impediments, reflects the complex and multifaceted nature of AI development in medicine. This demonstrates that although there is a general recognition of the importance of ethics, the extent to which it is prioritized differs significantly among experts. This diversity highlights the challenges in balancing ethical considerations with other developmental goals, such as performance optimization, economic viability, and technological innovation.

The analysis of the expert interviews identified 3 critical themes: first, the incompleteness of data and the far-reaching consequences associated with it; second, the renunciation of ethical requirements because of economic pressure; and third, the opinion that adhering to ethical standards would stand in the way of technological progress. These themes, reflecting a spectrum of perspectives from foundational importance to perceived obstacles, are explored in detail in subsequent sections, providing a deeper understanding of the multifaceted nature of ethics in AI development for medicine.

Incompleteness of Data

Quote A.4 (section *Subcategory 2: Consequences*) refers to the relevance of biases in the data. The lack of representativeness of the data, which underlies the development of AI-based applications, has been cited as a fundamental potential bias. Although awareness of the potential consequences, such as discrimination against certain population groups, is a crucial first step, it is not enough to merely recognize the issue to avoid potentially significant consequences [38]. Therefore, active measures must be taken to prevent these biases and ensure that AI-based applications do not perpetuate or exacerbate inequalities, thereby limiting potential harm.

To mitigate bias risks, developers should adopt comprehensive strategies, such as inclusive data collection methods, algorithmic audits, thorough testing across various demographic groups, and ongoing bias monitoring throughout the AI application lifecycle. As highlighted in quote A.1, the anticipatory ethical question in AI development is “extremely difficult,” underscoring the complexity of ensuring that AI systems are ethically sound and free from biases that could lead to discrimination or harm. Interdisciplinary teams, including ethicists and representatives from diverse communities, should guide the development process to ensure that ethical considerations are at the forefront of AI development.

A potential consequence of nonrepresentative data, as highlighted in quote A.4, includes “false negatives” in medicine, which are test results that incorrectly turn out to be negative despite the presence of diagnostic features of the disease under investigation [25]. However, it is also critical to recognize that

the same issue of nonrepresentativeness can lead to “false positives,” where tests incorrectly indicate the presence of a condition that is actually absent [25]. Both types of diagnostic inaccuracies have serious implications for patient care and treatment outcomes. This is further compounded by the sentiment expressed in quote A.3, where the need for everyone working with AI, especially in medicine, to consider the potential consequences of their work is emphasized, indicating a broader responsibility beyond development teams. This emphasizes the need for a comprehensive approach to diagnostic accuracy that accounts for both the presence of representative data and various factors influencing AI performance, extending beyond data representativeness [26]. Accuracy is also determined by the quality and variety of information subject to analysis from AI-based applications, including clinical, laboratory, and patient-reported data [39]. Furthermore, how AI processes and interprets this information, such as through its underlying algorithms and decision-making logic, is highly important for diagnostic accuracy [40]. There must be a match between the design purpose of the algorithm and real-world scenarios in which it is applied.

Moreover, the diagnostic accuracy of AI-based applications depends substantially on the proficiency with which health care professionals use these tools and their capacity to interpret and act on AI-generated recommendations [41]. For instance, if AI applications are used beyond their original scope without proper recalibration or validation for new populations or diseases, there is a risk of introducing errors, including false negatives and false positives [25].

False negatives in a clinical context can lead to physicians feeling a false sense of security and the diseases of patients remaining untreated for a long time [25]. Conversely, false positives can result in unnecessary treatments when a test erroneously indicates the presence of a disease, leading to significant consequences, such as unwarranted radiation exposure [25]. The psychological impact on patients, resulting from both false negatives and false positives, is a further concern that merits attention because of its effect on patient well-being and trust in medical systems.

The ethical implications of AI development, particularly when personal data are used, are highlighted in quote A.5 (section *Subcategory 3: Data Protection*). The use of training data for diagnosing specific diseases requires a careful ethical approach, particularly to understand the personal and clinical contexts from which such data are derived. This is particularly important for diseases that restrict the ability of the affected individuals to provide informed consent. Furthermore, ongoing discussions within ethics working groups about ethical processing, as mentioned in quote A.5, play a crucial role in safeguarding the dignity and rights of individuals whose data are used in these systems. Therefore, developers must recognize the sensitivity of medical data and the need for ethical considerations to be integrated from the outset of AI development for medical applications. Such early integration of ethics serves not only to enhance the accuracy and reliability of AI tools but also to safeguard the dignity and rights of individuals whose data are used in these systems.

Economic Pressure

The quotes from the second main category “results in the foreground” suggest that although the interviewed experts are aware of the relevance of ethics in the development of AI-based applications, it is in conflict with their own or demanded result orientation. A possible reason for the experts’ assessment is mentioned in quote B.3 (section *Subcategory 2: Economic Efficiency*). The profitability of AI developing companies is cited as one of the reasons why ethics is subordinate to the results in practice. Companies’ economic success pressure is decisive for the success pressure of all the employees involved in development. This conflict is further illustrated in quote B.2, where an expert highlights the competitive nature of AI development, suggesting that there are “more important things than ethics” in the context of existing competition. This perspective underscores the challenge of balancing ethical considerations with the need for AI applications to perform well in competitive markets.

As quote B.1 (section *Subcategory 1: Performance*) illustrates, the best possible performance is the focus of the development. Ethics indirectly plays a role here; quote B.3 implies, in this sense, the possibility of crossing “ethical boundaries” in favor of profitability. In addition to the deliberate crossing of boundaries, this statement also implies the possibility of unconscious disregard for ethics in the development of AI-based applications. The subordinate role of ethics in profitability in development and the associated noncompliance with potential boundaries is particularly severe, as the field of application is medicine. The sentiment of economic pressure overshadowing ethical considerations is also echoed in quote B.4, in which an expert states the importance of focusing on creating a product that is financially well, often at the expense of being ethically correct.

In addition to the relevance of ethics in relation to the use of human data and the potential consequences of a lack of representativeness, patient safety should always be at the center of the development of medical products and technologies. An excessive focus on the profitability of an application can lead to the marketing of immature or faulty products, which threaten patient welfare. Furthermore, as highlighted in quote B.3, the pursuit of profitability can sometimes lead developers to overlooking ethical boundaries, potentially resulting in products that have not been thoroughly evaluated for ethical considerations and patient safety. In addition to a direct threat to patient welfare and safety, a high susceptibility to error can also lead to rejection by users and a potentially irretrievable loss of trust [42].

Obstacle to Progress

Although the second main category cites result orientation because of economic pressure as a reason for the subordination of ethics, the third main category summarizes statements that view ethics as an “obstacle to progress.” The statements of experts in this category clearly show a rejection of ethics because of various demands and boundaries that are perceived as obstacles to the development of AI-based applications. Although no specific reasons for this assessment are provided, based on the knowledge of the steps relevant to development,

it can be assumed that the statements primarily refer to regulations and requirements in the sense of a necessary positive vote by ethics committees. For data collection, use, or evaluation in the context of developing AI-based applications, compliance with certain boundaries and regulations is indispensable, not only in the medical context. However, this essential compliance is sometimes perceived by experts as a balancing act, where meeting ethical demands can create challenges in advancing AI technology (quote C.1).

These boundaries and regulations serve to protect the participants and their data. If patient data are to be used, a positive vote from an ethics committee that certifies the safety of patients and their data is necessary to begin with the respective research and data use. As ethics committees' decisions can be time intensive depending on the type of planned research or data use and often require corrections on the part of the applicants, it is assumed that the necessity of a positive vote is one of the reasons that is viewed as an obstacle to progress. Furthermore, as highlighted in quote C.2, frustration with ethics being viewed as a blockade is evident: "Please stop bothering me on the topic of ethics in AI. It blocks at all corners and edges," illustrating the tension between the desire for rapid AI development and the need for ethical oversight. Although it can be assumed that AI-based applications would be developed faster if no vote from an ethics committee was necessary and patient data could be used directly, the resulting consequences for patients and citizens (think of the insurance industry) at least require critical evaluation.

Furthermore, although the need for a positive vote by an ethics committee can be anticipated as a perceived obstacle to progress in the development of AI by experts, it is also important to consider ongoing regulatory efforts, such as the proposed "Artificial Intelligence Act" by the European Parliament [43]. This regulation aims to harmonize rules on AI across the European Union, focusing on human-centric and trustworthy AI. The Act emphasizes the protection of health, safety, fundamental rights, and environmental concerns from potential harm caused by AI systems. It includes specific recommendations for high-risk AI systems, such as AI-based applications for medicine, demanding transparency, accountability, and accuracy in AI applications, especially those that may significantly impact individuals' rights and safety. The Act further acknowledges the ethical considerations in AI development and underscores the need for AI systems to adhere to robust ethical and legal standards. The regulatory requirement to adhere to ethical standards, as mandated by the Act, could further reinforce the perception of ethics and regulations being an obstacle, highlighting the tension between rapid technological advancement and the need for responsible innovation. In addition, quote C.3 conveys a sentiment shared by some experts that although ethical considerations are undeniably important, they are sometimes viewed as hindrances to meaningful AI advancement, further highlighting the complex dynamics between ethical considerations and the pursuit of technological progress in AI.

Consequences of Neglecting Ethics in the Development of AI-Based Applications in Medicine

Overview

If ethics is not considered in the development process of AI-based applications, it can have far-reaching consequences for patients and physicians, such as loss of trust and erosion of patient-centered care. This section focuses on the possible consequences of neglecting ethics when developing AI-based applications in medicine. In this context, the consequences for patients and likely main users (physicians) were considered.

Possible Consequences for Patients

If those responsible do not consider or only marginally consider the basic ethical principles in the development process of AI-based applications, various indirect and direct consequences can occur for the patients in whom the respective AI-based applications are used. The following examples illustrate the possible consequences of not considering ethical principles in the development process of AI in medicine:

- Misdiagnosis and diminished therapy outcomes: a lack of ethical considerations in the practical development process of AI-based applications can lead to biases in the training data used for development. For example, if the applications are used for diagnosis, the lack of representativeness of the data for certain population groups or individuals can lead to a higher susceptibility to errors. The results presented by AI can lead to potentially significant consequences for patients, such as overtreatment or undertreatment, resulting in diminished therapeutic outcomes, particularly in the absence of control by users [11]. These errors, stemming from a misdiagnosis because of unrepresentative data, challenge the principle of justice by threatening equitable medical care and contravene the principle of nonmaleficence by risking patient harm through inappropriate medical procedures [24]. Moreover, susceptibility to errors may directly compromise patient outcomes, especially when undertreatment occurs because of delayed or missed treatments from false-negative results [16]. The interrelated consequences of misdiagnosis and therapy outcomes highlight the critical need for user oversight and inclusion of diverse data sets in AI development to uphold ethical standards and patient care quality.
- Loss of trust: faulty diagnoses and the possibility of AI-based applications yielding discriminatory results can significantly undermine patient trust [44]. Such erosion of trust may lead patients to view AI-based medical applications skeptically, potentially refraining from using them in their treatment. This skepticism can hinder the integration of advanced AI tools in health care, which, if more accurate than physicians' assessments, could otherwise enhance patient outcomes. A loss of trust not only impedes technological adoption but can also indirectly challenge the principle of care, which is dedicated to optimizing patient welfare. Furthermore, patient reluctance to embrace AI solutions may inadvertently perpetuate inequalities in health care, particularly if AI facilitates more effective and efficient clinical practice. The reluctance to use AI technologies could result in disparity in care quality, as

physicians may be limited in their capabilities without AI support, ultimately affecting the standard of care provided. Moreover, an erosion or lack of trust in AI because of missing ethical oversight in development could extend to the physician-patient relationship and the overall health care sector. Moreover, an erosion or lack of trust in AI because of missing ethical oversight in development could extend to the physician-patient relationship and the overall health care sector [45]. This could lead to a general skepticism toward medical advice and a hesitation to participate in newer forms of treatment, potentially reverting to more traditional but less efficient methods. The physician-patient relationship is foundational to effective health care, as it relies on mutual trust and the belief that the best possible treatment options are being used, including ethically developed AI applications.

- Data misuse: a lack of consideration of ethics in the development of AI-based applications can lead to violations of existing data protection laws and misuse of patient data [46]. Patients who provide their data for research purposes and for the development of new applications in medicine must be able to rely on careful and legally compliant handling of their data, particularly in terms of informed consent and cybersecurity. Given the lack of traceability, informed consent is crucial, as patients must have a clear understanding of how their data will be used and the ability to consent to specific uses. This is of particular importance because health-related data include personal and sensitive information about patients. Ignoring existing regulations and ethical principles can result in highly sensitive patient data becoming accessible to companies, organizations, or individuals without consent [46]. This could have far-reaching consequences such as compromising patient privacy, enabling identity theft, or even affecting the broader integrity of medical research and public trust in the health care system. Similarly, robust cybersecurity measures are essential to protect sensitive health information from unauthorized access and breaches. Failure to implement such measures can lead to the exposure of personal health data, resulting in a loss of patient trust, potential harm, and a violation of the autonomy of patients if they lose control over their own data.
- Erosion of patient-centered care: the exclusion of patient values and preferences during the development of AI-based medical applications can have profound consequences. When AI systems are designed without a thorough understanding of patient autonomy, self-determination, and individual health goals, there is a risk of eroding the essence of patient-centered care [47]. AI recommendations that do not account for these personal factors might lead to a mechanical and less social approach to health care that could disregard the nuanced needs and desires of patients. For example, if AI tools are optimized solely for clinical efficiency without considering patient comfort and personal treatment preferences, they may suggest interventions that patients find unacceptable or intrusive. This misalignment can result in decreased adherence to treatment plans, loss of trust in the physician-patient relationship, and diminished health outcomes [48]. Given the importance of autonomy

in the physician-patient relationship and patient care in general, AI-based applications should be designed to support a shared decision-making model in which AI assists the therapeutic process rather than diminishing it. This would ensure that AI acts as an aid rather than a replacement for the human element in health care, empowering patients to be active participants in their treatment decisions rather than passive recipients of care.

Potential Consequences for Physicians

In addition to the significant consequences for patients, the lack of ethical consideration in the development process of AI-based applications in medicine can also lead to equally relevant impacts on anticipated primary users of the technology. Although the following examples primarily aim to illustrate the direct consequences for physicians, they also indirectly affect the patients being treated:

- Loss of credibility: potential errors in diagnosis or treatment recommendations resulting from inadequately trained AI applications can also significantly influence the societal image of the medical profession and its associated credibility [49]. Assuming that physicians continue to serve as the link between technology and patients, erroneous decisions based on the use of AI in medicine can be directly associated with the decision-making abilities of physicians, which can negatively impact their credibility and trust in the medical community [44]. Knowledge about the potential for discrimination of certain population groups by AI-based applications, which do not consider ethical guidelines in their development, can further shake patients' beliefs that physicians guarantee equal treatment for all. Because a patient's medical treatment often appears nontransparent and incomprehensible, the credibility of the medical community is an essential prerequisite for the physician-patient relationship [49].
- Rejection: the lack of consideration of ethics in the development of AI-based applications for use in a clinical context can lead to both indirect (eg, because of the consequences of incorrect diagnoses) and direct (eg, because of the lack of consideration of ethical principles) rejection of the technology by physicians. The rejection of AI-based applications can significantly impact the quality of medical care and the technological progress in medicine. Without the acceptance and trust of prospective primary users of the technology, the widespread use of AI-based applications in medicine is unlikely, as economic incentives for development are lacking. A rejecting attitude on the part of physicians can in this context also negatively impact future medical care quality considering the expected advantages of using AI in medicine [46].
- Legal consequences: the use of AI-based applications developed without considering ethical principles can lead to various legal consequences for users [50]. In addition to consequences based on state legislation and jurisprudence, professional legal consequences for physicians are also conceivable when using AI-based applications without considering ethical principles, as they form the basis of medical action. Besides the direct legal implications for physicians, health care organizations, such as hospitals,

clinics, or research institutions, may also be subject to significant responsibilities and potential liabilities when deploying AI-based applications that may not fully align with ethical and regulatory standards. In the case of erroneous AI decisions, which directly or indirectly result in diminished patient outcomes, the question of legal liability often remains unanswered [51]. As AI-based applications in medicine are likely to continue to be used and developed in a supportive role, it is assumed that the final decision-making and treatment recommendations will remain the responsibility of physicians. Thus, physicians not only act as a link between technologies and patients but also play a central role in adhering to ethical principles in medical care. Against this background, the use of AI-based applications in medicine developed without considering ethics can have legal consequences for both developers and users. In addition to the legal consequences of erroneous medical treatments, the use of AI-based applications without considering ethical principles also raises questions regarding the liability for violation of existing data protection and equal treatment laws [51]. In particular, failure to comply with data protection laws can compound these legal issues. Violations of patients' privacy rights through the mishandling of sensitive patient data, whether because of inadequate security measures, hacks, or unauthorized data sharing, may subject various entities, such as hospitals, clinics, research institutions, AI technology developers, and users to significant legal liability [50]. These data breaches not only compromise patient confidentiality but also could lead to a risk of regulatory sanctions for the involved entities, including substantial fines and potentially the loss of professional licenses. Therefore, AI development processes should incorporate robust data protection protocols to prevent legal repercussions and consequences for both patients and physicians. Adherence to ethical and legal standards should not merely be a regulatory requirement but a fundamental component of responsible and trustworthy health care innovation, vital for maintaining the integrity of patient care and the broader medical profession.

Limitations

This study's exploration of expert perspectives on ethics in AI development for medical applications, although insightful, encounters several limitations that are important to acknowledge. First, the geographical focus of the study was confined to Germany, potentially limiting the applicability of its findings to a global context in which cultural, legal, and ethical norms may vary. The selection of experts, although experienced in the development of AI-based applications in medicine, represents a relatively small and specific segment of the broader field. Moreover, the focus of the study, predominantly on experts with technical backgrounds in the development of AI-based applications, may lead to a narrowed perspective, given the lack of input from ethical professionals. Furthermore, the subjective nature of expert interviews should be considered because the responses are influenced by each expert's personal experiences and potential biases, which may not comprehensively represent the spectrum of views in the field.

Methodologically, the study's qualitative approach and reliance on secondary analysis of expert interviews inherently limits the generalizability of the results. Interpretations may be influenced by the research team's perspectives, and certain nuances in experts' statements may be overlooked. Although this study presents a secondary analysis of existing data, it is important to recognize the possibility of confirmation and selection bias during the initial data collection phase. The research methodology used could have unintentionally emphasized certain themes or perspectives, potentially aligning with the original researchers' preconceived notions or expectations. In addition, because of the limited number of experts included in the analysis and incomplete data saturation in some subcategories, certain aspects may not have been fully explored.

Furthermore, the findings of this study reflect a specific point in time in a rapidly evolving field. Therefore, the perspectives and opinions of experts may change as new developments, regulations, and ethical guidelines emerge. Although substantial, the focus on the development of AI-based applications in medicine does not encompass the entire spectrum of AI applications within the health care sector, excluding administrative and operational uses. Language and translation limitations may also have affected the study, as the original German interviews were translated into the English language. The subtle nuances of language and cultural context might be lost or misinterpreted in this translation process.

To address these limitations and enrich future research in this area, it is recommended that subsequent studies incorporate a broader and more diverse pool of experts, including professionals from ethical, legal, and patient advocacy backgrounds. Expanding the geographical scope to include experts from various cultural and legal contexts would also provide a global perspective on the ethical implications of developing AI-based applications for medicine. Methodologically, integrating both qualitative and quantitative approaches could offer a more comprehensive view, although ongoing research is required, considering the rapid advancements in AI and evolving ethical standards. By expanding the scope and methodology of future studies, a more nuanced and representative exploration of the ethical landscape of AI development for medical applications can be achieved.

Summary and Outlook

This study explored the importance of ethics in the development of AI-based medical applications by analyzing interviews with experts in the field of AI development. There was substantial variance in the assessment of the importance of ethics in the development of the AI-based applications. Although some of the interviewed experts classified ethics as an essential basis for development, others focused on good performance or economic efficiency. The results of the qualitative analysis also suggest that ethics is seen by some experts as an obstacle to progress, implying that it will be given little importance in the further development of AI-based applications. In addition to the subsequent discussion of the content analysis results, a particular focus was placed on the consequences that could arise from the lack of ethical considerations in the development of AI-based applications in medicine.

Although the results do not allow for generalization, because of the number of interviewees and the selected qualitative research method not meeting representative demands, the statements of the interviewed experts should be seen as an essential basis for further research and discussions because of recurring motives and new insights. A lack of ethical considerations in the development of AI-based applications can have significant consequences for patients. In addition to the danger of misconduct (eg, because of a lack of representativeness of the data sets used for development), a lack of consideration of ethical principles in the development of AI-based applications can also lead to a loss of trust from patients and potentially diminished therapy outcomes. When considering the possible impacts on physicians, the lack of consideration of ethics in the development process can lead to loss of credibility and rejection of technology.

Owing to technological progress in the field of AI, further reinforced, for example, by the development and broad availability of AI-based chat applications such as ChatGPT, there has been ongoing effort to develop guidelines and laws to guide the development and use of AI. Although such regulatory efforts, such as the “Artificial Intelligence Act” for harmonized rules on AI from the European Parliament, aim to provide a comprehensive regulatory framework and guideline for the development and use of AI, there is ongoing criticism and discussion about the adequacy and effectiveness of these guidelines in the rapidly evolving field of AI. In this context, it is important to emphasize that the sole availability of guidelines and laws does not ensure compliance. Therefore, although guidelines and laws are important to guide the development and use of AI, especially in the field of medicine, and when dealing with sensitive patient data, more work needs to be done to ensure compliance.

Moreover, the question arises as to whether mere adherence to these guidelines and laws is sufficient for the development of

ethical AI. Guidelines often provide a baseline for legal compliance, but ethical AI development demands a deeper and more nuanced understanding and application of ethical principles. Ethical AI goes beyond legal requirements to encompass ethical principles, such as respect for autonomy or justice in its algorithms, data handling, and decision-making processes. This requires continuous ethical assessment and reflection throughout the lifecycle of AI-based applications, from development to deployment, and beyond. Consequently, although following established guidelines is an important step in the development of AI, it is not the endpoint. Developers and users of AI-based applications in medicine need to engage in an ongoing dialog with diverse stakeholders such as ethicists, patients, and the broader community to anticipate, identify, and address emerging ethical challenges. This approach ensures that the development of AI is not just about complying with regulations but is intrinsically driven by a commitment to ethical responsibility and the betterment of patient care.

Furthermore, possible reasons for noncompliance with potential guidelines and low prioritization of ethics, such as the need for economic efficiency, should be critically examined. This includes assessing perspectives that view ethics as an obstacle to progress, as noted by some participating experts. Such critical evaluation is vital for ensuring the ethical development of AI-based applications, particularly in the field of medicine. Ethical considerations are fundamental to every approval process for AI-based applications to ensure the best possible and equal medical care for patients. Therefore, physicians should critically question the use of AI-based applications in the clinical context. In this regard, there needs to be a sufficient availability of opportunities to acquire further competencies to promote an understanding of technology and the related relevance of ethics. Only in this manner can the safety and best possible treatment of patients be ensured, as well as medical and technological progress, through AI.

Conflicts of Interest

None declared.

References

1. Hamet P, Tremblay J. Artificial intelligence in medicine. *Metabolism* 2017 Apr;69S:S36-S40. [doi: [10.1016/j.metabol.2017.01.011](https://doi.org/10.1016/j.metabol.2017.01.011)] [Medline: [28126242](https://pubmed.ncbi.nlm.nih.gov/28126242/)]
2. Eysenbach G. The role of ChatGPT, generative language models, and artificial intelligence in medical education: a conversation with ChatGPT and a call for papers. *JMIR Med Educ* 2023 Mar 06;9:e46885 [FREE Full text] [doi: [10.2196/46885](https://doi.org/10.2196/46885)] [Medline: [36863937](https://pubmed.ncbi.nlm.nih.gov/36863937/)]
3. McCarthy J, Minsky ML, Rochester N, Shannon CE. A Proposal for the Dartmouth Summer Research Project on artificial intelligence. *AI Magazine* 1955 Aug 31;27(4):12 [FREE Full text]
4. What is artificial intelligence and how is it used? European Parliament. 2020 Sep 4. URL: <https://www.europarl.europa.eu/news/en/headlines/society/20200827STO85804/what-is-artificial-intelligence-and-how-is-it-used> [accessed 2023-11-19]
5. McCarthy J. What is artificial intelligence? Stanford University. 2007 Nov 12. URL: <http://www-formal.stanford.edu/jmc/whatisai.pdf> [accessed 2023-11-19]
6. Russell S, Norvig P. *Artificial Intelligence: A Modern Approach*. London, UK: Pearson Education; Feb 16, 2010.
7. Zhou L, Sordo M. Expert systems in medicine. In: *Artificial Intelligence in Medicine: Technical Basis and Clinical Applications*. Cambridge, UK: Academic Press; 2021.
8. Kononenko I. Machine learning for medical diagnosis: history, state of the art and perspective. *Artif Intell Med* 2001 Aug;23(1):89-109. [doi: [10.1016/s0933-3657\(01\)00077-x](https://doi.org/10.1016/s0933-3657(01)00077-x)] [Medline: [11470218](https://pubmed.ncbi.nlm.nih.gov/11470218/)]

9. Bishop CM. Pattern Recognition and Machine Learning All: "Just the Facts 101" Material. Cham, Switzerland: Springer; 2006.
10. Doshi RH, Bajaj SS, Krumholz HM. ChatGPT: temptations of progress. *Am J Bioeth* 2023 Apr;23(4):6-8. [doi: [10.1080/15265161.2023.2180110](https://doi.org/10.1080/15265161.2023.2180110)] [Medline: [36853242](https://pubmed.ncbi.nlm.nih.gov/36853242/)]
11. Scerri M, Grech V. Artificial intelligence in medicine. *Early Hum Dev* 2020 Jun;145:105017. [doi: [10.1016/j.earlhumdev.2020.105017](https://doi.org/10.1016/j.earlhumdev.2020.105017)] [Medline: [32201033](https://pubmed.ncbi.nlm.nih.gov/32201033/)]
12. Hickman SE, Woitek R, Le EP, Im YR, Mouritsen Luxhøj C, Aviles-Rivero AI, et al. Machine learning for workflow applications in screening mammography: systematic review and meta-analysis. *Radiology* 2022 Jan;302(1):88-104 [FREE Full text] [doi: [10.1148/radiol.2021210391](https://doi.org/10.1148/radiol.2021210391)] [Medline: [34665034](https://pubmed.ncbi.nlm.nih.gov/34665034/)]
13. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* 2017 Feb 02;542(7639):115-118 [FREE Full text] [doi: [10.1038/nature21056](https://doi.org/10.1038/nature21056)] [Medline: [28117445](https://pubmed.ncbi.nlm.nih.gov/28117445/)]
14. Xie Y, Jiang B, Gong E, Li Y, Zhu G, Michel P, et al. Journal club: use of gradient boosting machine learning to predict patient outcome in acute ischemic stroke on the basis of imaging, demographic, and clinical information. *AJR Am J Roentgenol* 2019 Jan;212(1):44-51. [doi: [10.2214/AJR.18.20260](https://doi.org/10.2214/AJR.18.20260)] [Medline: [30354266](https://pubmed.ncbi.nlm.nih.gov/30354266/)]
15. Musen MA, Middleton B, Greenes RA. Clinical decision-support systems. In: Shortliffe E, Cimino J, editors. *Biomedical Informatics*. London, UK: Springer; 2014.
16. Sutton RT, Pincock D, Baumgart DC, Sadowski DC, Fedorak RN, Kroeker KI. An overview of clinical decision support systems: benefits, risks, and strategies for success. *NPJ Digit Med* 2020 Feb 6;3:17 [FREE Full text] [doi: [10.1038/s41746-020-0221-y](https://doi.org/10.1038/s41746-020-0221-y)] [Medline: [32047862](https://pubmed.ncbi.nlm.nih.gov/32047862/)]
17. Schork NJ. Artificial intelligence and personalized medicine. *Cancer Treat Res* 2019;178:265-283 [FREE Full text] [doi: [10.1007/978-3-030-16391-4_11](https://doi.org/10.1007/978-3-030-16391-4_11)] [Medline: [31209850](https://pubmed.ncbi.nlm.nih.gov/31209850/)]
18. Tripathi MK, Nath A, Singh TP, Ethayathulla AS, Kaur P. Evolving scenario of big data and artificial intelligence (AI) in drug discovery. *Mol Divers* 2021 Aug 23;25(3):1439-1460 [FREE Full text] [doi: [10.1007/s11030-021-10256-w](https://doi.org/10.1007/s11030-021-10256-w)] [Medline: [34159484](https://pubmed.ncbi.nlm.nih.gov/34159484/)]
19. Morley J, Machado CC, Burr C, Cowls J, Joshi I, Taddeo M, et al. The ethics of AI in health care: a mapping review. *Soc Sci Med* 2020 Sep;260:113172. [doi: [10.1016/j.socscimed.2020.113172](https://doi.org/10.1016/j.socscimed.2020.113172)] [Medline: [32702587](https://pubmed.ncbi.nlm.nih.gov/32702587/)]
20. Weidener L, Fischer M. Teaching AI ethics in medical education: a scoping review of current literature and practices. *Perspect Med Educ* 2023;12(1):399-410 [FREE Full text] [doi: [10.5334/pme.954](https://doi.org/10.5334/pme.954)] [Medline: [37868075](https://pubmed.ncbi.nlm.nih.gov/37868075/)]
21. Leslie D. Understanding artificial intelligence ethics and safety: a guide for the responsible design and implementation of AI systems in the public sector. The Alan Turing Institute. 2019. URL: https://www.turing.ac.uk/sites/default/files/2019-06/understanding_artificial_intelligence_ethics_and_safety.pdf [accessed 2023-12-19]
22. Beauchamp TL, Childress JF. *Principles of Biomedical Ethics*. New York, NY: Oxford University Press; 2013.
23. Hagendorff T. The ethics of AI ethics: an evaluation of guidelines. *Minds Mach* 2020 Feb 01;30(1):99-120. [doi: [10.1007/s11023-020-09517-8](https://doi.org/10.1007/s11023-020-09517-8)]
24. Keskinbora KH. Medical ethics considerations on artificial intelligence. *J Clin Neurosci* 2019 Jun;64:277-282. [doi: [10.1016/j.jocn.2019.03.001](https://doi.org/10.1016/j.jocn.2019.03.001)] [Medline: [30878282](https://pubmed.ncbi.nlm.nih.gov/30878282/)]
25. Challen R, Denny J, Pitt M, Gompels L, Edwards T, Tsaneva-Atanasova K. Artificial intelligence, bias and clinical safety. *BMJ Qual Saf* 2019 Mar 12;28(3):231-237 [FREE Full text] [doi: [10.1136/bmjqs-2018-008370](https://doi.org/10.1136/bmjqs-2018-008370)] [Medline: [30636200](https://pubmed.ncbi.nlm.nih.gov/30636200/)]
26. Cirillo D, Catuara-Solarz S, Morey C, Guney E, Subirats L, Mellino S, et al. Sex and gender differences and biases in artificial intelligence for biomedicine and healthcare. *NPJ Digit Med* 2020 Jun 01;3(1):81 [FREE Full text] [doi: [10.1038/s41746-020-0288-5](https://doi.org/10.1038/s41746-020-0288-5)] [Medline: [32529043](https://pubmed.ncbi.nlm.nih.gov/32529043/)]
27. Daneshjou R, Smith MP, Sun MD, Rotemberg V, Zou J. Lack of transparency and potential bias in artificial intelligence data sets and algorithms: a scoping review. *JAMA Dermatol* 2021 Nov 01;157(11):1362-1369 [FREE Full text] [doi: [10.1001/jamadermatol.2021.3129](https://doi.org/10.1001/jamadermatol.2021.3129)] [Medline: [34550305](https://pubmed.ncbi.nlm.nih.gov/34550305/)]
28. Hedlund M, Persson E. Expert responsibility in AI development. *AI Soc* 2022 Jun 13. [doi: [10.1007/s00146-022-01498-9](https://doi.org/10.1007/s00146-022-01498-9)]
29. Pant A, Hoda R, Tantithamthavorn C, Turhan B. Ethics in AI through the developer's view: a grounded theory literature review. *arXiv Preprint* posted online June 20, 2022. [FREE Full text] [doi: [10.48550/arXiv.2206.09514](https://doi.org/10.48550/arXiv.2206.09514)]
30. Jobin A, Ienca M, Vayena E. The global landscape of AI ethics guidelines. *Nat Mach Intell* 2019 Sep 02;1(9):389-399. [doi: [10.1038/s42256-019-0088-2](https://doi.org/10.1038/s42256-019-0088-2)]
31. Mittelstadt B. Principles alone cannot guarantee ethical AI. *Nat Mach Intell* 2019 Nov 04;1(11):501-507. [doi: [10.1038/s42256-019-0114-4](https://doi.org/10.1038/s42256-019-0114-4)]
32. Sanderson C, Douglas D, Lu Q, Schleiger E, Whittle J, Lacey J, et al. AI ethics principles in practice: perspectives of designers and developers. *IEEE Trans Technol Soc* 2023 Jun;4(2):171-187. [doi: [10.1109/tts.2023.3257303](https://doi.org/10.1109/tts.2023.3257303)]
33. Morley J, Kinsey L, Elhalal A, Garcia F, Ziosi M, Floridi L. Operationalising AI ethics: barriers, enablers and next steps. *AI Soc* 2021 Nov 15;38(1):411-423. [doi: [10.1007/s00146-021-01308-8](https://doi.org/10.1007/s00146-021-01308-8)]
34. Weidener L, Fischer M. Artificial intelligence teaching as part of medical education: qualitative analysis of expert interviews. *JMIR Med Educ* 2023 Apr 24;9:e46428 [FREE Full text] [doi: [10.2196/46428](https://doi.org/10.2196/46428)] [Medline: [36946094](https://pubmed.ncbi.nlm.nih.gov/36946094/)]

35. O'Brien BC, Harris IB, Beckman TJ, Reed DA, Cook DA. Standards for reporting qualitative research: a synthesis of recommendations. *Acad Med* 2014 Sep;89(9):1245-1251 [FREE Full text] [doi: [10.1097/ACM.0000000000000388](https://doi.org/10.1097/ACM.0000000000000388)] [Medline: [24979285](https://pubmed.ncbi.nlm.nih.gov/24979285/)]
36. Dresing T, Pehl T. *Praxisbuch Interview, Transkription & Analyse: Anleitungen und Regelsysteme für qualitativ Forschende*. Hessen, Germany: Dr. Dresing und Pehl; 2015.
37. Mayring P. *Qualitative Content Analysis: A Step-by-Step Guide*. Thousand Oaks, CA: SAGE Publications; 2021.
38. McLennan S, Fiske A, Tigard D, Müller R, Haddadin S, Buyx A. Embedded ethics: a proposal for integrating ethics into the development of medical AI. *BMC Med Ethics* 2022 Jan 26;23(1):6 [FREE Full text] [doi: [10.1186/s12910-022-00746-3](https://doi.org/10.1186/s12910-022-00746-3)] [Medline: [35081955](https://pubmed.ncbi.nlm.nih.gov/35081955/)]
39. Jayakumar S, Sounderajah V, Normahani P, Harling L, Markar SR, Ashrafian H, et al. Quality assessment standards in artificial intelligence diagnostic accuracy systematic reviews: a meta-research study. *NPJ Digit Med* 2022 Jan 27;5(1):11 [FREE Full text] [doi: [10.1038/s41746-021-00544-y](https://doi.org/10.1038/s41746-021-00544-y)] [Medline: [35087178](https://pubmed.ncbi.nlm.nih.gov/35087178/)]
40. Feng J, Phillips RV, Malenica I, Bishara A, Hubbard AE, Celi LA, et al. Clinical artificial intelligence quality improvement: towards continual monitoring and updating of AI algorithms in healthcare. *NPJ Digit Med* 2022 May 31;5(1):66 [FREE Full text] [doi: [10.1038/s41746-022-00611-y](https://doi.org/10.1038/s41746-022-00611-y)] [Medline: [35641814](https://pubmed.ncbi.nlm.nih.gov/35641814/)]
41. Lambert SI, Madi M, Sopka S, Lenes A, Stange H, Buszello CP, et al. An integrative review on the acceptance of artificial intelligence among healthcare professionals in hospitals. *NPJ Digit Med* 2023 Jun 10;6(1):111 [FREE Full text] [doi: [10.1038/s41746-023-00852-5](https://doi.org/10.1038/s41746-023-00852-5)] [Medline: [37301946](https://pubmed.ncbi.nlm.nih.gov/37301946/)]
42. Lockey S, Gillespie N, Holm D, Someh IA. A review of trust in artificial intelligence: challenges, vulnerabilities and future directions. In: *Proceedings of the 54th Hawaii International Conference on System Sciences*. 2021 Presented at: 54th Hawaii International Conference on System Sciences; January 5, 2021; Kauai, Hawaii URL: <https://scholarspace.manoa.hawaii.edu/items/f7c9bf5a-19fa-4a9e-a3cc-114ec8b529e4> [doi: [10.24251/hicss.2021.664](https://doi.org/10.24251/hicss.2021.664)]
43. Artificial intelligence act. European Parliament. URL: https://www.europarl.europa.eu/doceo/document/TA-9-2023-0236_EN.pdf [accessed 2023-11-19]
44. Hatherley JJ. Limits of trust in medical AI. *J Med Ethics* 2020 Jul 27;46(7):478-481. [doi: [10.1136/medethics-2019-105935](https://doi.org/10.1136/medethics-2019-105935)] [Medline: [32220870](https://pubmed.ncbi.nlm.nih.gov/32220870/)]
45. Gille F, Jobin A, Ienca M. What we talk about when we talk about trust: theory of trust for AI in healthcare. *Intell Based Med* 2020 Nov;1-2:100001. [doi: [10.1016/j.ibmed.2020.100001](https://doi.org/10.1016/j.ibmed.2020.100001)]
46. Price WN2, Cohen IG. Privacy in the age of medical big data. *Nat Med* 2019 Jan 7;25(1):37-43 [FREE Full text] [doi: [10.1038/s41591-018-0272-7](https://doi.org/10.1038/s41591-018-0272-7)] [Medline: [30617331](https://pubmed.ncbi.nlm.nih.gov/30617331/)]
47. Bjerring JC, Busch J. Artificial intelligence and patient-centered decision-making. *Philos Technol* 2020 Jan 08;34(2):349-371. [doi: [10.1007/s13347-019-00391-6](https://doi.org/10.1007/s13347-019-00391-6)]
48. Sauerbrei A, Kerasidou A, Lucivero F, Hallowell N. The impact of artificial intelligence on the person-centred, doctor-patient relationship: some problems and solutions. *BMC Med Inform Decis Mak* 2023 Apr 20;23(1):73 [FREE Full text] [doi: [10.1186/s12911-023-02162-y](https://doi.org/10.1186/s12911-023-02162-y)] [Medline: [37081503](https://pubmed.ncbi.nlm.nih.gov/37081503/)]
49. Ahuja AS. The impact of artificial intelligence in medicine on the future role of the physician. *PeerJ* 2019 Oct 4;7:e7702 [FREE Full text] [doi: [10.7717/peerj.7702](https://doi.org/10.7717/peerj.7702)] [Medline: [31592346](https://pubmed.ncbi.nlm.nih.gov/31592346/)]
50. Price WN2, Gerke S, Cohen IG. Potential liability for physicians using artificial intelligence. *JAMA* 2019 Nov 12;322(18):1765-1766. [doi: [10.1001/jama.2019.15064](https://doi.org/10.1001/jama.2019.15064)] [Medline: [31584609](https://pubmed.ncbi.nlm.nih.gov/31584609/)]
51. Schneeberger D, Stöger K, Holzinger A. The European legal framework for medical AI. In: *Proceedings of the International Cross-Domain Conference for Machine Learning and Knowledge Extraction*. 2020 Presented at: International Cross-Domain Conference for Machine Learning and Knowledge Extraction; August 25-28, 2020; Dublin, Ireland. [doi: [10.1007/978-3-030-57321-8_12](https://doi.org/10.1007/978-3-030-57321-8_12)]

Abbreviations

- AI:** artificial intelligence
CDSS: Clinical Decision Support Systems
ML: machine learning
-

Edited by K El Emam, B Malin; submitted 24.07.23; peer-reviewed by A Marušić, G Lorenzini, D Chrimes; comments to author 28.10.23; revised version received 20.11.23; accepted 09.12.23; published 12.01.24.

Please cite as:

Weidener L, Fischer M

Role of Ethics in Developing AI-Based Applications in Medicine: Insights From Expert Interviews and Discussion of Implications
JMIR AI 2024;3:e51204

URL: <https://ai.jmir.org/2024/1/e51204>

doi: [10.2196/51204](https://doi.org/10.2196/51204)

PMID: [38875585](https://pubmed.ncbi.nlm.nih.gov/38875585/)

©Lukas Weidener, Michael Fischer. Originally published in JMIR AI (<https://ai.jmir.org>), 12.01.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR AI, is properly cited. The complete bibliographic information, a link to the original publication on <https://www.ai.jmir.org/>, as well as this copyright and license information must be included.

Original Paper

Beyond the Hype—The Actual Role and Risks of AI in Today's Medical Practice: Comparative-Approach Study

Steffan Hansen¹, MA; Carl Joakim Brandt¹, PhD; Jens Søndergaard¹, PhD

Research Unit of General Practice, Institution of Public Health, University of Southern Denmark, Odense, Denmark

Corresponding Author:

Steffan Hansen, MA

Research Unit of General Practice

Institution of Public Health

University of Southern Denmark

J.B. Winsløvs Vej 9

Odense, 5000

Denmark

Phone: 45 65 50 36 19

Email: sholsthansen@health.sdu.dk

Abstract

Background: The evolution of artificial intelligence (AI) has significantly impacted various sectors, with health care witnessing some of its most groundbreaking contributions. Contemporary models, such as ChatGPT-4 and Microsoft Bing, have showcased capabilities beyond just generating text, aiding in complex tasks like literature searches and refining web-based queries.

Objective: This study explores a compelling query: can AI author an academic paper independently? Our assessment focuses on four core dimensions: relevance (to ensure that AI's response directly addresses the prompt), accuracy (to ascertain that AI's information is both factually correct and current), clarity (to examine AI's ability to present coherent and logical ideas), and tone and style (to evaluate whether AI can align with the formality expected in academic writings). Additionally, we will consider the ethical implications and practicality of integrating AI into academic writing.

Methods: To assess the capabilities of ChatGPT-4 and Microsoft Bing in the context of academic paper assistance in general practice, we used a systematic approach. ChatGPT-4, an advanced AI language model by Open AI, excels in generating human-like text and adapting responses based on user interactions, though it has a knowledge cut-off in September 2021. Microsoft Bing's AI chatbot facilitates user navigation on the Bing search engine, offering tailored search

Results: In terms of relevance, ChatGPT-4 delved deeply into AI's health care role, citing academic sources and discussing diverse applications and concerns, while Microsoft Bing provided a concise, less detailed overview. In terms of accuracy, ChatGPT-4 correctly cited 72% (23/32) of its peer-reviewed articles but included some nonexistent references. Microsoft Bing's accuracy stood at 46% (6/13), supplemented by relevant non-peer-reviewed articles. In terms of clarity, both models conveyed clear, coherent text. ChatGPT-4 was particularly adept at detailing technical concepts, while Microsoft Bing was more general. In terms of tone, both models maintained an academic tone, but ChatGPT-4 exhibited superior depth and breadth in content delivery.

Conclusions: Comparing ChatGPT-4 and Microsoft Bing for academic assistance revealed strengths and limitations. ChatGPT-4 excels in depth and relevance but falters in citation accuracy. Microsoft Bing is concise but lacks robust detail. Though both models have potential, neither can independently handle comprehensive academic tasks. As AI evolves, combining ChatGPT-4's depth with Microsoft Bing's up-to-date referencing could optimize academic support. Researchers should critically assess AI outputs to maintain academic credibility.

(JMIR AI 2024;3:e49082) doi:[10.2196/49082](https://doi.org/10.2196/49082)

KEYWORDS

AI; artificial intelligence; ChatGPT-4; Microsoft Bing; general practice; ChatGPT; chatbot; chatbots; writing; academic; academia; Bing

Introduction

Artificial intelligence's (AI) journey has been nothing short of incredible. Starting with its early days of rule-based systems, we have seen it grow and mature, stepping into the realm of machine learning, and more recently, diving into deep learning. This transformative journey has shaken up a lot of sectors, but health care is where AI has truly left an indelible mark.

Today, algorithms can spot issues in our x-rays or magnetic resonance imaging, sometimes even better than our seasoned doctors [1]. AI does not just stop there; it even gives us a heads-up on potential life-threatening situations in intensive care units, predicting conditions like septic shock hours before they occur. The world of drug discovery is moving faster than ever, thanks to AI's helping hand [2]. However, as with most things, there are issues. There are big questions about how we protect our data and ensure different health record systems talk to each other [3], not to mention the lingering worries about biases in AI and the sometimes uneasy feeling of trusting a machine we do not fully "get" [4].

When you look at the big picture, we see ground-breaking models like GPT-3, ChatGPT-4 [5,6], and Microsoft Bing [7] making waves. They are not just about churning out text. They are doing things we had never imagined, like assisting in literature searches or refining our everyday web-based searches [8]. Their accomplishments in challenges, such as the Turing Test [9] and the LAMBADA (LAnguage Modeling Broadened to Account for Discourse Aspects) tasks [10], just go on to show how capable they are. Comparing powerhouses like ChatGPT-4 and Bing is not just for fun; it gives us a glimpse into where AI's language abilities might be headed, and with new kids on the block like Google Bard, the sky is the limit [11]. Writing an academic paper, though? That is still a world where the human touch shines. From combing through mountains of literature to connecting the dots in innovative ways, it is a craft that demands the very best of us, but here is a thought: given how far AI has come, could it, one day, pen down an academic masterpiece on its own? This paper is all about that tantalizing question.

As we embark on this exploration, we will keenly assess a few critical dimensions:

- **Relevance:** can AI ensure that its response precisely addresses the prompt and brings to the table information that is truly pertinent to the question or topic?
- **Accuracy:** how reliable is AI in delivering information that is not just factually correct but also up-to-date with the current pulse of the academic field?
- **Clarity:** when we read what is written by AI, does it resonate with clarity, coherence, and a logical flow of ideas, all presented with precise and unambiguous language?
- **Tone and style:** given the seriousness of academic papers, can AI match the appropriate tone and style, ensuring it resonates with the formality and professionalism we expect to see in academic texts?

We are diving deep to see if AI can muster up the relevance, accuracy, clarity, and tone we associate with academic work,

and of course, while we probe these questions, we are not losing sight of the overarching ethics and practicality of inviting AI into the revered domain of academic writing.

Methods

Ethical Considerations

In Denmark, ethical committee approval is only mandatory for studies that include trials involving liveborn human individuals, human gametes intended for fertilization, fertilized human eggs, embryonic cells and embryos, tissue, cells and genetic material from humans, embryos, etc, or deceased persons. Also included are clinical trials of medicines in humans and clinical trials of medical devices. Hence, our study did not require approval from an ethical committee.

Overview

In this methods section, we have detailed the approach taken to evaluate and compare the performance of ChatGPT-4 and Microsoft Bing in the context of assisting with an academic paper in the realm of general practice. This section outlines the data collection process, prompt design, evaluation criteria, and analysis of the AI-generated responses.

Models

ChatGPT-4

ChatGPT-4 is an advanced AI language model developed by OpenAI [5], based on the ChatGPT-4 architecture. It is designed to generate human-like text and engage in interactive conversations with users. Trained on a vast data set, ChatGPT-4 demonstrates a strong understanding of context, language, and reasoning abilities. When using GPT-4, it is important to highlight that during a conversation, the information and discussion are dynamically shaped throughout the interaction. Indeed, GPT-4 can respond by incorporating the information the user provides, potentially leading to different outcomes even for users with similar queries. This dynamic nature is crucial for understanding how a large language model like GPT-4 operates.

Although ChatGPT-4 can perform various tasks, such as answering questions, providing recommendations, and generating content, it has a knowledge cut-off date of September 2021. This means that the model has been trained on a data set consisting of text and information available up until that point. Therefore, any events, advancements, or changes in various fields that have occurred since September 2021 will not be known to ChatGPT-4. Additionally, it should be noted that ChatGPT-4, like any AI language model, reflects the data on which it has been trained. As a result, its knowledge might contain inaccuracies, biases, or outdated information even for events and topics within its known time frame.

Microsoft Bing

The Microsoft Bing AI chatbot [7] is an intelligent conversational agent developed by Microsoft Corporation, designed to assist users in navigating the Microsoft Bing search engine and answering various queries. Leveraging AI, natural language processing, and machine learning, the Microsoft Bing

AI chatbot understands user inputs and provides relevant information or search results accordingly. Integrated seamlessly with the Microsoft Bings platform, the chatbot offers a user-friendly and interactive way to engage with search functionalities, enhancing the overall user experience.

Prompt Design

In the context of AI, especially with large language models, a “prompt” refers to a set of instructions or a question given to the AI to guide its response. The purpose of a prompt is to set clear expectations for the AI’s output and to ensure that the response generated aligns with the user’s intent.

A prompt was designed to secure the AI models’ ability to understand and generate accurate, relevant, and coherent responses in a formal and professional tone. Each prompt provided the AI models with the context of an academic paper and set the tone and expectations for the responses. The following specific prompt was used to ensure that both ChatGPT-4 and Microsoft Bing were primed for the task at hand:

I need your help with an academic paper. Please provide me with clear and concise explanations, using evidence and logical reasoning to support your responses. Your tone should be formal and professional, and your language should be free from errors and ambiguity. I am looking for accurate and well-supported information that will help me to achieve my academic goals.

Data Collection

The interview with the 2 models took place on March 9, 2023, with early access to ChatGPT-4. Both ChatGPT-4 and Microsoft Bing were asked to provide an outline for a discussion article on the chosen topic, encompassing various aspects of general practice. This approach aimed to evaluate the AI models’ ability to synthesize information and structure a coherent, well-organized outline that could serve as a foundation for a comprehensive discussion article. As differences between the outlines are likely, the most comprehensive outline was used to ensure a meaningful comparison between interviews. The length of each question was limited to ensure accuracy and reduce the risk of errors during the conversation.

Evaluation Criteria

It is important to note that the evaluation was conducted solely by one author, and the assessments were largely based on their subjective judgment. To compare and assess the quality of the AI-generated responses, the following evaluation criteria were established:

- **Relevance:** the extent to which the AI-generated response addresses the prompt and provides information pertinent to the question or topic.
- **Accuracy:** the degree to which the information provided is factually correct and up to date, based on the current state of knowledge in the field.

- **Clarity:** the clarity and coherence of the AI-generated response, including the logical flow of ideas and the use of precise, unambiguous language.
- **Tone and style:** the appropriateness of the tone and style of the AI-generated response, considering the formal and professional context of an academic paper.

To evaluate the evaluation criteria, a comprehensive literature search was conducted to identify areas where AI might be useful and implemented in general practice.

Analysis

Each AI-generated response was analyzed independently, using the evaluation criteria, providing the strengths and weaknesses of each model. Hereafter, a comparison between the 2 models was conducted to establish differences. The results of the evaluation and comparison between the 2 models were then compiled and analyzed to determine the overall performance of ChatGPT-4 and Microsoft Bing related to the area of AI use in general practice and the areas preidentified, aiming at identifying the strengths and weaknesses of each AI model as well as any potential areas for improvement.

Results

For a complete comparison, the full conversation with both ChatGPT-4 and Microsoft Bing models can be found in [Multimedia Appendix 1](#).

Relevance

Chat-GPT

GPT-4 offers a detailed analysis of AI applications in health care, focusing on general practice, its limitations, ethical concerns, and the importance of collaboration between AI and health care professionals. It provides comprehensive information, citing academic sources and studies, discussing AI algorithms, natural language processing, pattern recognition, evidence-based medicine, and personalized treatment plans. ChatGPT-4 also addresses data privacy, security concerns, and technical challenges while emphasizing the need to integrate AI systems with clinical workflows and patient needs. It provides a relevant and comprehensive examination of AI’s potential benefits and challenges in health care, emphasizing the need for integration with clinical workflows and a balanced approach to ensure optimal patient care.

Microsoft Bing

Microsoft Bing offers a brief overview of AI in general practice, addressing advantages and limitations without delving into specific applications or ethical considerations. It lacks the depth and citations and does not emphasize the importance of collaboration between AI and health care professionals. Although Microsoft Bing touches on themes that are relevant, it provides neither specific study references nor in-depth explanations, offering a more concise perspective ([Table 1](#)).

Table 1. Comparison of ChatGPT and Microsoft Bing in terms of topic relevance.

Evaluation criteria	ChatGPT-4	Microsoft Bing
Relevance	<ul style="list-style-type: none"> A detailed analysis of AIa applications in healthcare Comprehensive information and citing academic sources Emphasizing the need for integration with clinical workflows and a balanced approach to ensure optimal patient care 	<ul style="list-style-type: none"> A brief overview of AI in general practice Lack of in-depth or specific study citations Offering a more concise perspective

^aAI: artificial intelligence.

Accuracy

ChatGPT

ChatGPT-4 included 23 of 32 (72%) precise peer-reviewed articles with high accuracy. The introduction and applications in general practice were 100% correct. However, it also cited 9 nonexistent articles, with 4 out of 7 inaccuracies in limitations and all 4 ethical considerations being inaccurately cited.

Microsoft Bing

Microsoft Bing included 6 of 13 (46%) highly accurate, peer-reviewed articles, along with 7 non-peer-reviewed but highly relevant articles. Ethical considerations and applications in general practice cited 3 and 2 non-peer-reviewed articles, respectively (Table 2).

The references provided from both models, along with the accuracy distribution, can be found in [Multimedia Appendix 2](#).

Table 2. Comparison of ChatGPT and Microsoft Bing in terms of accuracy.

Evaluation criteria	ChatGPT-4	Microsoft Bing
Accuracy	<ul style="list-style-type: none"> A total of 23 out of 32 (72%) precise peer-reviewed articles, with high accuracy A total of 9 nonexistent articles, with specific inaccuracies 	<ul style="list-style-type: none"> A total of 6 out of 13 (46%) highly accurate, peer-reviewed articles A total of 7 non-peer-reviewed but highly relevant articles

Clarity

Chat GPT-4

Overall, the text generated by ChatGPT demonstrates a high level of clarity and coherence, exhibiting a logical flow of ideas and the use of precise, unambiguous language. The text is easy to follow and understand, even for readers who may not be familiar with the technical terms and concepts discussed.

Microsoft Bing

Similar to ChatGPT, the text exhibits a high level of clarity and coherence, with a logical flow of ideas and the use of precise, unambiguous language. It is easily comprehensible, even for readers unfamiliar with the technical terms and concepts discussed. However, the text could be improved by providing more details and examples to support the points made, as many areas are discussed in a more general manner (Table 3).

Table 3. Comparison of ChatGPT and Microsoft Bing in terms of clarity.

Evaluation Criteria	ChatGPT-4	Microsoft Bing
Clarity	The text is clear, coherent, and easy to understand, even for nontechnical readers.	The text is clear and coherent but could benefit from more detailed examples.

Tone (Chat GPT-4 and Microsoft Bing)

Overall, the tone and style of the text are appropriate for the formal and professional context of an academic paper,

effectively conveying complex ideas in a clear and objective manner (Table 4).

Table 4. Comparison of ChatGPT and Microsoft Bing in terms of tone.

Evaluation criteria	ChatGPT-4	Microsoft Bing
Tone	Appropriate for an academic paper, conveying ideas clearly and objectively	Appropriate for an academic paper, conveying ideas clearly and objectively

Discussion

Principal Findings

In recent years, AI has become an increasingly prevalent tool in various domains, including health care and academic research. AI language models, such as ChatGPT-4 and Microsoft Bing, have demonstrated the potential to assist researchers in

generating and organizing content for academic papers. In the context of general practice, a rapidly evolving field with a growing need for accurate and relevant information, understanding the strengths and limitations of these AI models is crucial for researchers and practitioners alike. This paper aimed to compare and analyze the performance of ChatGPT-4 and Microsoft Bing in assisting with an academic paper in general practice, focusing on their relevance, accuracy, clarity,

as well as tone and style. By examining their respective contributions and limitations, we seek to provide insights into their potential uses and areas for improvement in AI-assisted research.

In terms of relevance, ChatGPT-4 provided a detailed analysis of AI applications in health care, emphasizing the importance of collaboration between AI and health care professionals, while Microsoft Bing offered a concise overview without delving into specific applications or ethical considerations. As for accuracy, ChatGPT-4 accurately cited 72% (23/32) of peer-reviewed articles, but it also inaccurately cited 9 nonexistent articles. Microsoft Bing, on the other hand, included 6 of 13 (46%) accurate peer-reviewed articles and 7 non-peer-reviewed but highly relevant articles.

Regarding clarity, both ChatGPT-4 and Microsoft Bing demonstrated high levels of clarity and coherence, presenting a logical flow of ideas with precise, unambiguous language. Nevertheless, Microsoft Bing could benefit from providing more details and examples to support its points, as certain areas were discussed in a more general manner. Lastly, in terms of tone and style, both AI models used an appropriate tone and style for the formal and professional context of an academic paper, effectively conveying complex ideas in a clear and objective manner.

Comparison With the Existing Literature

The results of this study, which compared the performance of ChatGPT-4 and Microsoft Bing in assisting with an academic paper in general practice, can be contextualized within the broader landscape of AI applications in health care and general practice research. The findings align with several previous studies that have highlighted the potential of AI language models, such as ChatGPT-4, to deliver relevant, detailed, and coherent information on complex subjects like health care [6,12].

The superior performance of ChatGPT-4 in providing comprehensive and in-depth analysis aligns with its advanced architecture and extensive training on a vast data set, which has been documented to enable the model to generate human-like text and engage in interactive conversations with users [12]. Similarly, the results are consistent with previous research that has emphasized the importance of collaboration between AI and health care professionals to achieve optimal patient care [13].

However, the observed weaknesses in ChatGPT-4's accuracy, specifically in citing nonexistent articles, highlight the limitations of AI language models in some areas of academic research. This issue has been acknowledged in existing

literature, where concerns have been raised about the potential for AI-generated content to include inaccuracies, biases, or misinformation [14].

In contrast, Microsoft Bing's more concise approach to providing information echoes its primary function as a search engine assistant rather than a specialized AI language model. This result is consistent with the notion that AI chatbots, while capable of providing relevant information, may not always deliver the depth and detail required for more demanding academic tasks [15].

Strengths

This study has some strengths, as follows:

- Prompt design: the study used a well-crafted prompt to ensure that both ChatGPT-4 and Microsoft Bing were primed for the task, which helped in generating accurate, relevant, and coherent responses in a formal and professional tone.
- Evaluation criteria: the established evaluation criteria (relevance, accuracy, clarity, as well as tone and style) provided a comprehensive framework for comparing and assessing the quality of the AI-generated responses.
- Analysis: the independent analysis of each AI-generated response, followed by a comparison between the 2 models, allowed for a thorough understanding of the strengths and weaknesses of each AI model.

Weaknesses

The weaknesses of the study are the following:

- Data collection: the study's data collection method, which involved interviewing the 2 models, may have been limited in scope. A more comprehensive approach involving a larger sample of questions or topics could have provided a broader understanding of the AI models' capabilities.
- Knowledge cut-off: ChatGPT-4 has a knowledge cut-off date of September 2021, which may have limited its ability to provide up-to-date information in some instances.
- Limited exploration of AI models: the study only compared 2 AI models—ChatGPT-4 and Microsoft Bing. This may not provide a complete picture of the landscape of AI tools available for assisting with academic papers in general practice. Including more AI models, such as Google's chatbot—Bard, in the comparison could have yielded a more comprehensive analysis. However, this model is not currently available in Denmark.

The strengths and weaknesses of each model are presented in Table 5.

Table 5. A side-by-side comparison of the features and aspects of ChatGPT-4 and Microsoft Bing's artificial intelligence (AI) chatbot.

Feature or aspect	ChatGPT-4	Microsoft Bing
Developer	OpenAI	Microsoft Corporation
Primary function	Generating human-like text and engaging in interactive conversations	Assisting users in navigating the Microsoft Bing search engine and answering queries
Training or technology	Vast data set, context understanding, language, and reasoning abilities	Artificial intelligence, natural language processing, and machine learning
Special features	Answering questions, providing recommendations, and generating content	Integrating with the Bing platform and enhancing the search experience
Conversation limits	25 conversations per 3 hours	Limited to 20 prompts
Internet access	No	Yes
Knowledge cut-off	Up to 2021	Uses OpenAI technology with access to the internet and thus can acquire the newest information
Memory constraints	Forgets information within longer conversations and might stop midsentence in lengthy responses	Closely related to ChatGPT-4 in this area
Additional information	Some responses may require user prompts to be complete	Offers a user-friendly and interactive way to engage with search functionalities

Implications for AI-Assisted Research

The findings of this study have several implications for researchers and practitioners using AI in general practice and other academic fields. These implications are as follows:

- **Quality of AI-generated content:** the comparison between ChatGPT-4 and Microsoft Bing demonstrates that the quality of AI-generated content can vary between models. Researchers and practitioners should be aware of the strengths and weaknesses of different AI models when selecting a tool to assist with their work.
- **Importance of collaboration:** both ChatGPT-4 and Microsoft Bing highlight the importance of collaboration between AI and health care professionals. AI systems should be designed to complement human expertise and foster collaboration, enhancing the overall quality of research and practice.
- **Relevance and accuracy:** ensuring the relevance and accuracy of AI-generated responses is crucial for researchers and practitioners. Although AI models can provide valuable insights, they might also generate inaccuracies or outdated information. Users must verify the information provided by AI models and cross-check it with up-to-date, reliable sources.
- **Clarity and tone:** AI-generated content should be clear and coherent; it should maintain an appropriate tone and style for the intended audience. Although AI models like ChatGPT-4 and Microsoft Bing show promising results in these aspects, users should carefully review and edit the generated content to ensure it meets the required standards.
- **Ethical considerations:** as AI continues to be integrated into various aspects of research and practice, ethical considerations must be addressed. Data privacy, security, and responsible use of AI-generated content are crucial to ensuring that AI is used responsibly and effectively in general practice and other academic fields.

Overall, the findings of this study indicate that AI models, such as ChatGPT-4 and Microsoft Bing, can provide valuable assistance in general practice and other academic fields. However, researchers and practitioners should be aware of the limitations and potential pitfalls of AI-generated content and use these tools thoughtfully and responsibly.

Areas for Improvement and Future Research

AI Model Improvements

ChatGPT-4

Although ChatGPT-4 demonstrates strong performance in relevance, clarity, and tone, there is room for improvement in terms of accuracy, especially in relation to citing nonexistent articles. Enhancing the fact-checking and source validation capabilities of the model could help address this issue.

Microsoft Bing

Microsoft Bing could benefit from improvements in providing more in-depth, relevant content with proper citations. Enhancing the model's understanding of specific academic contexts and ethical considerations would allow it to provide more comprehensive and valuable insights to the users.

Methodology Improvements

The methodology improvements required are as follows:

- **Expanding the sample size:** including more AI models in the comparison would provide a broader understanding of the capabilities and limitations of AI-assisted research.
- **Diversifying the topics:** evaluating AI-generated responses across a wider range of topics and academic fields could offer more generalizable insights into the strengths and weaknesses of AI-assisted research.
- **Including human evaluation:** adding a panel of human evaluators to assess the AI-generated content could help provide a more nuanced understanding of the quality and relevance of the responses.

Future Research Directions

Some directions for future research are explained below:

- Longitudinal studies: investigating the evolution of AI models over time, as they are updated and trained on new data, could provide valuable insights into the progress of AI-assisted research and the potential of these tools in various academic fields.
- Ethical implications: examining the ethical implications of AI-generated content in academic research, such as issues related to plagiarism, data privacy, and potential biases, could help develop best practices and guidelines for responsible use of AI in research.
- Integration with research workflows: exploring how AI models can be effectively integrated into existing research workflows and practices and identifying the most effective ways to combine AI-generated content with human expertise would help maximize the benefits of AI-assisted research.

By addressing these areas for improvement and exploring future research directions, researchers and practitioners can continue to refine the use of AI models in general practice and other academic fields, ultimately enhancing the quality, efficiency, and impact of their work.

Conclusions

Our study comparing ChatGPT-4 and Microsoft Bing in assisting with writing an academic paper in general practice yielded several key findings. ChatGPT-4 demonstrated strong performance in terms of relevance, clarity, and tone, providing comprehensive information and detailed analysis of AI applications in health care. However, it exhibited weaknesses in accuracy, particularly in citing nonexistent articles. Microsoft Bing offered a more concise perspective, touching on relevant themes but lacking depth and proper citations.

In terms of methods used, the study incorporated prompt design, data collection, evaluation criteria, and analysis of AI-generated

responses. The strengths of these methods include the design of a prompt that effectively engaged both AI models and the establishment of clear evaluation criteria. However, there is room for improvement in the methodology, such as expanding the sample size, diversifying the topics, and including human evaluation.

When comparing ChatGPT-4 and Microsoft Bing, ChatGPT-4 emerged as a more capable AI model for assisting with an academic paper in general practice. It provided a more in-depth, relevant, and coherent analysis of the topic; however, improvements in accuracy, particularly in source validation, would further enhance its utility. On the other hand, Microsoft Bing could benefit from improvements in providing more comprehensive content and proper citations to better support academic research.

In conclusion, ChatGPT-4 and Microsoft Bing present distinct pros and cons in academic writing. ChatGPT-4 excels in relevance and depth, but both AI models require improvement. Merging their strengths can produce comprehensive answers from ChatGPT-4 and up-to-date references from Microsoft Bing.

Despite their impressive abilities, these tools currently cannot author articles independently in certain areas. As AI models advance and incorporate current references and critical thinking, they may eventually conduct and create research autonomously.

This study's findings hold substantial implications for AI-assisted research across diverse fields, emphasizing areas for refinement and future research directions to optimize AI models in academia. To mitigate risks, researchers must adopt a critical approach, corroborate information from various sources, and stay aware of AI models' limitations. This approach allows them to harness AI while preserving the integrity and rigor of their work.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Interview with models.

[[DOCX File , 50 KB - ai_v3i1e49082_app1.docx](#)]

Multimedia Appendix 2

References provided by models and their relevance.

[[DOCX File , 22 KB - ai_v3i1e49082_app2.docx](#)]

References

1. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* 2017 Feb 02;542(7639):115-118 [[FREE Full text](#)] [doi: [10.1038/nature21056](https://doi.org/10.1038/nature21056)] [Medline: [28117445](https://pubmed.ncbi.nlm.nih.gov/28117445/)]
2. Blanco-González A, Cabezón A, Seco-González A, Conde-Torres D, Antelo-Riveiro P, Piñeiro Á, et al. The role of AI in drug discovery: challenges, opportunities, and strategies. *Pharmaceuticals (Basel)* 2023 Jun 18;16(6):891 [[FREE Full text](#)] [doi: [10.3390/ph16060891](https://doi.org/10.3390/ph16060891)] [Medline: [37375838](https://pubmed.ncbi.nlm.nih.gov/37375838/)]

3. Raghupathi W, Raghupathi V. Big data analytics in healthcare: promise and potential. *Health Inf Sci Syst* 2014 Feb 7;2(1):3 [FREE Full text] [doi: [10.1186/2047-2501-2-3](https://doi.org/10.1186/2047-2501-2-3)] [Medline: [25825667](https://pubmed.ncbi.nlm.nih.gov/25825667/)]
4. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* 2019 Oct 25;366(6464):447-453. [doi: [10.1126/science.aax2342](https://doi.org/10.1126/science.aax2342)] [Medline: [31649194](https://pubmed.ncbi.nlm.nih.gov/31649194/)]
5. GPT-4 is OpenAI's most advanced system, producing safer and more useful responses. OpenAI. URL: <https://openai.com/product/gpt-4> [accessed 2023-11-01]
6. Brown T, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P. Language models are few-shot Learners. arXiv. Preprint posted online on May 28, 2020 [FREE Full text] [doi: [10.48550/arXiv.2005.14165](https://doi.org/10.48550/arXiv.2005.14165)]
7. Introducing the new Bing. Bing. URL: <https://www.bing.com/new> [accessed 2023-11-01]
8. Radford A, Wu J, Child R, Luan D, Amodei D, Sutskever I. Language Models are Unsupervised Multitask Learners. URL: https://d4mucfpsywv.cloudfront.net/better-language-models/language_models_are_unsupervised_multitask_learners.pdf [accessed 2023-11-01]
9. Turing AM. I.—Computing machinery and intelligence. *Oxford Academic* 1950:433-460. [doi: [10.1093/mind/LIX.236.433](https://doi.org/10.1093/mind/LIX.236.433)]
10. Paperno D, Kruszewski G, Lazaridou A, Pham Q, Bernardi R, Pezzelle S. The LAMBADA dataset: Word prediction requiring a broad discourse context. arXiv. Preprint posted online on Jun 20, 2016 [FREE Full text] [doi: [10.18653/v1/p16-1144](https://doi.org/10.18653/v1/p16-1144)]
11. Booth A, Papaioannou D, Sutton A. *Systematic Approaches to a Successful Literature Review*. Thousand Oaks, CA: Sage; 2012.
12. Radford A, Narasimhan K, Salimans T, Sutskever I. Improving Language Understanding by Generative Pre-Training. 2018. URL: https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language-unsupervised/language_understanding_paper.pdf [accessed 2023-11-27]
13. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. *Nat Med* 2019 Jan 7;25(1):44-56. [doi: [10.1038/s41591-018-0300-7](https://doi.org/10.1038/s41591-018-0300-7)] [Medline: [30617339](https://pubmed.ncbi.nlm.nih.gov/30617339/)]
14. Bender E, Gebru T, McMillan-Major A, Shmitchell S. On the dangers of stochastic parrots: can language models be too big? In: *FACCT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. Virtual Event Canada; 2021 Presented at: FACCT '21; March 3-10; Virtual event Canada URL: <https://dl.acm.org/doi/10.1145/3442188.3445922> [doi: [10.1145/3442188.3445922](https://doi.org/10.1145/3442188.3445922)]
15. Brandtzaeg P, Følstad A. *Why People Use Chatbots*. Cham, Switzerland: Springer Link; 2017:377-392.

Abbreviations

AI: artificial intelligence

LAMBADA: LAngeuage Modeling Broadened to Account for Discourse Aspects

Edited by H Liu; submitted 17.05.23; peer-reviewed by M Salvagno, G Sebastian; comments to author 07.09.23; revised version received 11.10.23; accepted 15.10.23; published 22.01.24.

Please cite as:

Hansen S, Brandt CJ, Søndergaard J

Beyond the Hype—The Actual Role and Risks of AI in Today's Medical Practice: Comparative-Approach Study

JMIR AI 2024;3:e49082

URL: <https://ai.jmir.org/2024/1/e49082>

doi: [10.2196/49082](https://doi.org/10.2196/49082)

PMID:

©Steffan Hansen, Carl Joakim Brandt, Jens Søndergaard. Originally published in JMIR AI (<https://ai.jmir.org>), 22.01.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR AI, is properly cited. The complete bibliographic information, a link to the original publication on <https://www.ai.jmir.org/>, as well as this copyright and license information must be included.

Original Paper

An Environmental Uncertainty Perception Framework for Misinformation Detection and Spread Prediction in the COVID-19 Pandemic: Artificial Intelligence Approach

Jiahui Lu^{1,2}, PhD; Huibin Zhang², BE; Yi Xiao², PhD; Yingyu Wang², BE

¹State Key Laboratory of Communication Content Cognition, People's Daily Online, Beijing, China

²School of New Media and Communication, Tianjin University, Tianjin, China

Corresponding Author:

Huibin Zhang, BE

School of New Media and Communication

Tianjin University

Number 92, Weijin Road

Tianjin, 300072

China

Phone: 86 15135154977

Email: zhanghb@tju.edu.cn

Abstract

Background: Amidst the COVID-19 pandemic, misinformation on social media has posed significant threats to public health. Detecting and predicting the spread of misinformation are crucial for mitigating its adverse effects. However, prevailing frameworks for these tasks have predominantly focused on post-level signals of misinformation, neglecting features of the broader information environment where misinformation originates and proliferates.

Objective: This study aims to create a novel framework that integrates the uncertainty of the information environment into misinformation features, with the goal of enhancing the model's accuracy in tasks such as misinformation detection and predicting the scale of dissemination. The objective is to provide better support for online governance efforts during health crises.

Methods: In this study, we embraced uncertainty features within the information environment and introduced a novel Environmental Uncertainty Perception (EUP) framework for the detection of misinformation and the prediction of its spread on social media. The framework encompasses uncertainty at 4 scales of the information environment: physical environment, macro-media environment, micro-communicative environment, and message framing. We assessed the effectiveness of the EUP using real-world COVID-19 misinformation data sets.

Results: The experimental results demonstrated that the EUP alone achieved notably good performance, with detection accuracy at 0.753 and prediction accuracy at 0.71. These results were comparable to state-of-the-art baseline models such as bidirectional long short-term memory (BiLSTM; detection accuracy 0.733 and prediction accuracy 0.707) and bidirectional encoder representations from transformers (BERT; detection accuracy 0.755 and prediction accuracy 0.728). Additionally, when the baseline models collaborated with the EUP, they exhibited improved accuracy by an average of 1.98% for the misinformation detection and 2.4% for spread-prediction tasks. On unbalanced data sets, the EUP yielded relative improvements of 21.5% and 5.7% in macro-F1-score and area under the curve, respectively.

Conclusions: This study makes a significant contribution to the literature by recognizing uncertainty features within information environments as a crucial factor for improving misinformation detection and spread-prediction algorithms during the pandemic. The research elaborates on the complexities of uncertain information environments for misinformation across 4 distinct scales, including the physical environment, macro-media environment, micro-communicative environment, and message framing. The findings underscore the effectiveness of incorporating uncertainty into misinformation detection and spread prediction, providing an interdisciplinary and easily implementable framework for the field.

(JMIR AI 2024;3:e47240) doi:[10.2196/47240](https://doi.org/10.2196/47240)

KEYWORDS

misinformation detection; misinformation spread prediction; uncertainty; COVID-19; information environment

Introduction

Background

The World Health Organization and the United Nations have issued warnings about an “infodemic,” highlighting the spread of misinformation alongside the COVID-19 pandemic on social media [1]. Misinformation is characterized as “factually incorrect information not backed up by evidence” [2]. This misleading information frequently encompasses harmful health advice, misinterpretations of government control measures and emerging sciences, and conspiracy theories [3]. This phenomenon has inflicted detrimental impacts on public health, carrying “severe consequences with regard to people’s quality of life and even their risk of mortality” [4].

Automatic algorithms are increasingly recognized as valuable tools in mitigating the harm caused by misinformation. These techniques can rapidly identify misinformation, predict its spread, and have demonstrated commendable performance. The state-of-the-art detection techniques exhibit accuracy ranging from 65% to 90% [5,6], while spread-prediction techniques achieve performance levels between 62.5% and 77.21% [7,8]. The high accuracy of these techniques can be largely attributed to the incorporation of handcrafted or deep-learned linguistic and social features associated with misinformation [9-11]. Scholars have consistently invested efforts in integrating theoretically relevant features into algorithmic frameworks to enhance accuracy further.

Scholars have introduced diverse frameworks for misinformation detection and spread-prediction algorithms. Nevertheless, existing frameworks have predominantly concentrated on the intricate post-level signals of misinformation, emphasizing linguistic and social features (such as user relationships, replies, and knowledge sources) associated with misinformation. Notably, these frameworks have often overlooked the characteristics of the information environment in which misinformation originates and proliferates [12]. This neglect could potentially result in diminished performance for misinformation detectors when applied in various real-world misinformation contexts. This is due to the fact that different misinformation contexts possess unique characteristics within their information environment, influencing the types of misinformation that can emerge and thrive [13]. An indispensable characteristic of the information environment concerning misinformation is uncertainty. Uncertainty arises when the details of situations are ambiguous, complex, unpredictable, or probabilistic, and when information is either unavailable or inconsistent [14]. In uncertain situations, individuals tend to generate and disseminate misinformation as a means of resisting uncertainty and seeking understanding amid chaotic circumstances [15,16]. The COVID-19 pandemic serves as a notable example, marked by a lack of understanding of emerging science [17], uncertainties surrounding official guidelines and news reports [18], and unknown impacts on individuals and society [19]. Hence, in this study, we recognize uncertainty as the pivotal feature in the information environment of misinformation. Our objective is to formulate a novel framework for perceiving environmental uncertainty,

specifically tailored for the detection and spread prediction of misinformation during the COVID-19 pandemic.

Our contributions can be outlined as follows. Theoretically, we provide a comprehensive exploration of uncertainty across 4 distinct scales of the information environment, namely, the physical environment, macro-media environment, micro-communicative environment, and message framing. These scales collectively contribute to the emergence and dissemination of misinformation. Furthermore, we hold the distinction of being the pioneers in integrating Environmental Uncertainty Perception (EUP) into the realms of misinformation detection and spread prediction. In terms of methodology, we introduce the EUP framework, designed to capture uncertainty signals from the information environment of a given post for both misinformation detection and spread prediction. Our experiments conducted on real-life data underscore the effectiveness of the EUP framework.

This paper unfolds as follows: In the “Related Work” section, we provide a concise review of the related work. The “Proposed Theoretical Framework” section elucidates uncertainty features within the information environment, which are pertinent to misinformation detection and spread prediction. Moving on to the “Research Objectives” section, we outline our study objectives. The “Methods” section details our methodology for testing the proposed framework. In the “Data Set and Experiment” section, we present our data set, experiments, and comprehensive analyses. The “Discussion” section delves into discussions on our findings, unraveling the theoretical and practical implications of our work. Finally, the “Conclusions” section concludes with a summary and outlines directions for future research.

Related Work

Detecting misinformation on social media represents a burgeoning research field that has garnered considerable academic attention. Multiple frameworks have been put forth for this task, primarily falling into 2 approaches: the post-only approach and the “zoom-in” approach [12]. In the former, frameworks focus on studying post features to differentiate misinformation from general information. Linguistic features, including novelty, complexity, emotions, and content topics, are frequently explored [6,11]. Additionally, researchers have delved into multimodal features, particularly those based on visuals [20,21]. Deep learning models in natural language processing have also proven beneficial for the misinformation detection task [5,22].

The “zoom-in” approach places emphasis on socio-contextual signals, centering on users’ networking aspects (eg, user relationships, number of replies, number of created threads; [23,24]) and network characteristics (eg, degree centrality [25]). Another line of research underscores the significance of relevant knowledge sources, including fact-checking websites [26] and knowledge graphs [27], which can be used to validate specific claims of interest.

Recently, Sheng et al [12] introduced a “zoom-out” approach, concentrating on the information environments of misinformation that can offer signals for detection. In their

approach, they incorporated the news environment into fake news detection. Their hypothesis posited that fake news should not only be relevant but also novel and distinct from recent popular news, enabling them to capture audience attention and achieve widespread dissemination. Their findings revealed that signals of popularity and novelty can enhance the performance of state-of-the-art misinformation detectors.

In the realm of misinformation detection, misinformation spread prediction represents another challenging task, albeit one that has received limited attention. This task involves predicting whether a piece of misinformation is likely to be disseminated to a broader audience through actions such as likes, comments, and shares. Within this context, our specific focus is on predicting whether misinformation is likely to be retweeted. This can be viewed as a binary classification task, akin to misinformation detection. Frameworks for this task typically incorporate linguistic and social features, which may overlap with or differ from those used in misinformation detection. Linguistic features such as persuasive styles, emotional expressions, and message coherence prove valuable in predicting the spread of misinformation [28,29]. Additionally, social features, including user metadata (eg, number of friends, verification) and tweet metadata (eg, presence of images and URLs), are identified as relevant factors for predicting misinformation spread [25].

Proposed Theoretical Framework

Uncertainty as a Central Aspect in Misinformation

Our study builds upon Sheng et al's [12] "zoom-out" approach, adopting an interdisciplinary perspective that centers on the uncertainty within the information environment of misinformation. The realms of communication and psychology literature have conceptualized uncertainty as a fundamental aspect of misinformation. Uncertainty is said to prevail "when details of situations are ambiguous, complex, unpredictable, or probabilistic; uncertainty is also present when information is unavailable or inconsistent, and when individuals feel insecure about their own state of knowledge or the general state of knowledge" [14]. Confronted with uncertainty, individuals are driven to alleviate it by constructing their understanding of the situation [16]. This constructive process is known as sensemaking, which encompasses how individuals impart meaning to their surroundings and use it as a foundation for subsequent interpretation and action [30]. Sensemaking entails the utilization of information by individuals to fill gaps in their understanding [31]. Yet, the utilization of information in this manner does not always guarantee truth. In situations where information is slow to emerge, individuals are driven to comprehend uncertain situations by relying on their existing knowledge and heuristics for judgment. Unfortunately, this process often leads to the formation of false beliefs and misinformation [32]. Additionally, individuals may "turn to unofficial sources to satisfy their information needs," potentially exposing themselves to inaccurate information [33]. As suggested by Kim et al [34], exposure to misinformation has the potential to diminish feelings of uncertainty. Moreover, as individuals integrate more information into their comprehension of a situation, there is a tendency to seek plausibility, which

may lead to the generation and acceptance of misinformation [16,35].

The aforementioned tendencies are notably prominent in the context of the COVID-19 pandemic, as the pandemic represents a time of heightened uncertainty. The emergence of the pandemic was marked by a mysterious disease with previously unseen symptoms. Fundamental questions regarding the origins of the disease, measures for self-protection, and strategies for containing the outbreak were not immediately evident. As the pandemic progressed, uncertainty persisted regarding how and when the outbreak would be fully contained, as well as the long-term impact it would have on individuals and society. The uncertainty stemming from the pandemic, coupled with the surge of social media as a primary source of information, has facilitated the spread of misinformation [16].

Although many studies have identified "uncertainty" as a central aspect of misinformation, they have not thoroughly elucidated how uncertainty, as a crucial feature of the information environment, can aid in the detection of misinformation and the prediction of its spread. The literature frequently treats uncertainty as a static and holistic feature of a situation. However, the level of uncertainty within a situation can be dynamic, evolving as the situation progresses. For instance, uncertainties about the virus and the initial life changes induced by the COVID-19 pandemic would have been considerably higher at its onset than they are at present [36]. Moreover, uncertainty can manifest differently across various scales of the information environment. The information environment has become increasingly intricate with the proliferation of the internet and communication technologies. Individuals may be exposed to a substantial volume of information about trending topics through mainstream mass media (eg, newspapers, TV, social media trends) within a short time frame, constituting a macro-media environment. Simultaneously, they may selectively engage in detailed communications on a specific issue provided by self-media (eg, subscription accounts, self-broadcasting), shaping a micro-communicative environment. Uncertainty manifested in these 2 environments may independently or interactively influence people's sensemaking processes and, consequently, their outputs (eg, misinformation). Additionally, uncertainty can be inherent in the misinformation itself, providing cues for its detection and spread prediction. We will elaborate on the features of uncertainty in the information environment in the following section.

Uncertainty in the Information Environment

Uncertainty in the Physical Environment

Uncertainty prevails in the physical environment when unknown risks pose potential threats to our societal systems [15,16]. Scholars refer to such threats as "crises," which can encompass natural disasters, large-scale accidents, social security incidents, and public health emergencies such as the pandemic [37]. Crises are marked by the existence of uncertainty and the imperative for timely decision-making [38]. Therefore, a crucial process during crises is sensemaking. However, the efforts needed for sensemaking will vary as a crisis progresses through stages. The Crisis and Emergency Risk Communication Model delineates 5 common stages in the crisis life cycle, spanning

“from risk, to eruption, to clean-up and recovery, and on into evaluation [38].” The eruption of the crisis, also known as the breakout stage, occurs when a key event triggers the crisis [39]. This is the period when the public becomes initially aware of the crisis, characterized by mysteries and heightened motivation to make sense of it. Evidence indicates that the breakout stage of a crisis harbors the highest level of uncertainty and demands extensive sensemaking efforts (eg, government updates [40]; social media communication [41]), consequently leading to a higher incidence of misinformation [42]. This evidence implies that misinformation is more likely to surface and proliferate in tandem with uncertainty in the information environment during the breakout stage compared with other stages throughout a crisis. These insights offer valuable cues for the detection and prediction of misinformation during the COVID-19 pandemic.

Uncertainty in the Macro-Media Environment

The macro-media environment encompasses recent media opinions and public attention to trending topics [12]. Governments and mainstream media play a pivotal role in setting the agenda for public attention. During crises such as the COVID-19 pandemic, governments frequently make swift and crucial decisions to safeguard the public. However, these decisions are often made without sufficient transparency, leading to potential uncertainties surrounding their rationale [43]. Such decisions inevitably draw media and public attention, quickly becoming trending topics in mainstream media outlets [44,45]. Regrettably, these rapid decisions often leave audiences with a high level of uncertainty about the reasons behind and the processes involved in making these decisions, potentially paving the way for misinformation. Supporting this notion, Lu [3] identified a correlation between the swift decision to quarantine Wuhan city and the emergence of misinformation regarding government control measures during the early stages of the COVID-19 pandemic in China. The evidence presented indicates that when public attention is directed toward a trending topic that carries uncertainty, misinformation is likely to emerge and spread. In simpler terms, it can be anticipated that when a piece of information is associated with a trending topic characterized by high uncertainty (as opposed to low uncertainty), there is a higher probability that the information could be misinformation and disseminated.

Uncertainty in the Micro-Communicative Environment

Differing from the macro-media environment, which offers a macro perspective on what mass audiences have recently read and focused on, the micro-communicative environment provides a micro view of the communication surrounding a specific issue. Both media and individuals tend to communicate using frames or terms imbued with uncertainty when discussing matters that lack evidence or consensus, such as those stemming from emerging science during the COVID-19 pandemic [32,46]. As an illustration, in the initial phase of the pandemic, when Hong Kong officials reported the first instance of a dog testing “weakly positive” for COVID-19 infection, subsequent media reports highlighted that “Hong Kong scientists aren’t sure [emphasis added] if the dog is actually infected or if it picked up the virus from a contaminated surface [47].” Experimental evidence has shown that such uncertainty frames about scientific matters can diminish people’s trust in science [48]. Empirical

evidence from real-life social media data further indicates that a communication style marked by ambiguity can potentially lead audiences to generate and disseminate misinformation [32]. This body of findings implies that if information is embedded in uncertain (as opposed to consensus) communication, it is more likely to be misinformation and disseminated.

Uncertainty in Message Framing

Uncertainty can also manifest within the message through its framing or word choice. Uncertainty frames are prevalent in misinformation [15,49]. Oh et al [15] illustrated that source ambiguity and content ambiguity are 2 significant features of misinformation. When individuals create a piece of misinformation that lacks evidence and credibility, they often use uncertain words to describe the unreliable source (eg, someone) or the potential rationale (eg, possible, likely) behind the statement. The incorporation of uncertain words can indeed facilitate the spread of misinformation [29,50]. The inclusion of uncertainty expressions in messages leads individuals to perceive the information as more relevant and suitable for themselves [51]. Consequently, if misinformation exhibits a higher level of uncertainty, it is more likely to be accepted and disseminated by the public.

Research Objectives

Our research objective is to explore whether uncertainty features within the information environment can enhance the effectiveness of misinformation detection and spread prediction. To achieve this, we introduce a novel EUP framework specifically designed for both tasks. We seek to assess the standalone effectiveness of the EUP and anticipate that it can augment the capabilities of existing state-of-the-art misinformation detectors and predictors. Therefore, we conducted experiments to answer the following research questions:

- *Research question 1:* Can EUP be effective in misinformation detection and spread prediction?
- *Research question 2:* Can EUP improve the performances of the state-of-the-art algorithms for misinformation detection and spread prediction?

Methods

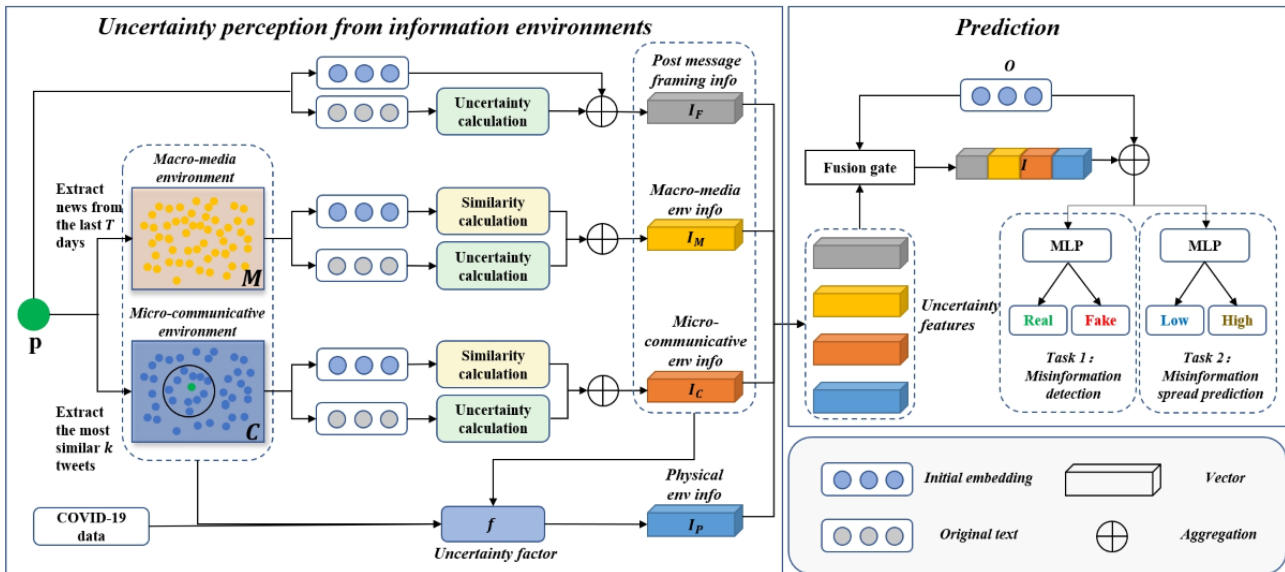
Overview

Figure 1 offers an overview of the EUP pipeline. The model consists of 4 uncertainty extraction components. Upon receiving a post (denoted as p), the initial step involves constructing its macro-media environment and micro-communicative environment. This is accomplished by extracting recent news and social media data, respectively. Subsequently, we use a probabilistic model and a similarity calculation method to derive the uncertainty information for the 2 environments mentioned above, denoted as I_M and I_C . Likewise, we utilized the probabilistic model to capture the uncertainty of the post p itself, resulting in the representation of message framing denoted as I_F . Simultaneously, the operationalization of uncertainty in the physical environment entails using the number of COVID-19 cases and the volume of news as key indicators, denoted as I_p .

Lastly, the 4 vectors are integrated using a gate guided by the extracted post feature o (which may not necessarily equal p) from the misinformation detector, such as bidirectional encoder representations from transformers (BERT) [52]. The fused

vectors I and o are then input into the final classifier, typically a multilayer perceptron (MLP), to predict whether p is fake or real in task 1 and low or high in task 2.

Figure 1. An environmental uncertainty perception (EUP) framework for misinformation detection and spread prediction in the COVID-19 pandemic.



Uncertainty Detection Model

For detecting uncertainty in natural language [53], we used a probabilistic model that considers the local n-gram features of sentences. Each n-gram is assigned a weight that reflects its tendency to convey uncertainty. The definition of each feature involves a quadruplet (type, size, context, and aggregation). “Type” signifies the type of n-gram considered, such as lemma or morphosyntactic pattern. “Size” indicates the size of the n-gram. “Context” serves as an indicator, specifying whether the weight is based on the occurrence frequency of the n-gram in an uncertain sentence or on the occurrence frequency of the n-gram as an uncertainty marker. “Aggregation” refers to the method used to consolidate different scores of the n-grams within a sentence. Multimedia Appendix 1 [49,54-57] furnishes a summary of the diverse features, denoted as F_i , that are scrutinized in the uncertainty detection model.

Next, we exemplify the calculation of uncertainty using 1 of these features, F_1 , as an illustration. F_1 is defined by the quadruplet (Lemma, 1, uncertainty marker, and sum). For each lemma w , we can compute the number of occurrences in the corpus, the number of occurrences in uncertain sentences, and the number of occurrences as an uncertainty marker, denoted as F_s , F_w , and F_m , respectively. The conditional probability of a lemma w becoming an uncertainty marker is calculated using the following equation:

$$p(c|w)=F_m/F_s \quad (1)$$

where c represents the class of context uncertainty under analysis, specifically whether it pertains to being an uncertainty marker. Additionally, we introduce a confidence score linked to the probability of mitigating the impact of instances where certain lemmas occur infrequently in the corpus yet yield a high probability:

$$\text{conf}(w)=1-(1-F_s) \quad (2)$$

F_1 takes into account both the conditional probability of each lemma w and the corresponding confidence score in the sentence s , and the formula is calculated as follows:

$$F_1 = \text{Mean}(\text{Norm}(\{F_i(s)\}_{i=1}^{|F|})) \quad (3)$$

Similarly, other features F_i can be derived using the above method. We generated the uncertainty of the whole sentence by mean pooling to represent the average uncertainty signals of F_i :

$$F^{A,\text{Mean}}(s)=\text{Mean}(\text{Norm}(\{F_i(s)\}_{i=1}^{|F|})) \quad (4)$$

where $\text{Norm}(\cdot)$ denotes the normalization.

Representation of the Macro-Media Environment

We collect news reports from mainstream media outlets released within T days before the post p is published to construct a macro-media environment according to the following definition:

$$M = \{e: e \in E, 0 \leq t_p - t_e \leq T\} \quad (5)$$

where E denotes the set of all collected news items, M denotes the set of news items in the macro-media environment of the post p , and t_p and t_e represent the release time of post p and news e , respectively. For post p or each news item e , the initial representations are the output of a pretrained language model (eg, BERT [52]), denoted as p and e , respectively.

The macro-media environment is expected to reflect the impact of a trending topic with high uncertainty on the veracity of a post. That is, if a post is related to a trending topic with (vs without) high uncertainty, it is then expected to be more likely misinformation and disseminated. To this end, the representation of the macro-media environment should consider both the correlation between the post and the environment and the

uncertainty of the environment. We first calculate cosine similarity between p and each news item e in E :

$$S(p,e) = (p \cdot e) / (|p| \cdot |e|) \quad (6)$$

We combine the similarity and environment representations to represent the similarity representation of a post p to the environment:



where e^M_i represents each news item in M and \boxtimes is the Hadamard product operator.

We then measure the uncertainty of the macro-media environment using the model described in the ‘‘Uncertainty Detection Model’’ section. The uncertainty representation of the macro-media environment, denoted as U_M , can be expressed by the following equation:



Finally, the macro-media environment of a post p is represented as an aggregation of the similarity representation of p to the environment (S_M) and the uncertainty representation of the environment (U_M) using an MLP, denoted as I_M :

$$I_M = \text{MLP}(S_M \boxtimes U_M) \quad (9)$$

where \boxtimes is the concatenation operator. The integration of an MLP is instrumental in the dual objective of retaining crucial information while concurrently achieving data dimensionality reduction. All MLPs are individually parameterized. We omit their index numbers in the above equations for brevity.

Representation of the Micro-Communicative Environment

We collected tweets from Twitter (X; X Corp.) published within T days before the post p was published to construct the micro-communicative environment. We calculated the similarity of all tweets to the post p and selected the top k of them, using them as a *micro-communicative environment* (C), which is defined as follows:

$$C' = \{v: v \in V, 0 \leq t_p - t_v \leq T\} \quad (10)$$

where V denotes the set of all collected tweet items and t_v represents the release time of the tweet v .

$$C = \{v: v \in \text{Topk}(p, C')\} \quad (11)$$

where $\text{Topk}(\cdot)$ represents the operation of selecting the k tweets that have the highest similarity to p , $k = r \cdot |C'|$, and $r \in (0,1)$ represents the percentage of extraction.

Using the same approach as in the previous 2 sections, we derive the similarity representation of the post p to the micro-communicative environment and the uncertainty representation of the environment:



Finally, the micro-communicative environment of a post p is represented as an aggregation of the similarity representation of a post p to the environment (S_C) and the uncertainty representation of the environment (U_C) using an MLP, denoted as I_C :

$$I_C = \text{MLP}(S_C U_C) \quad (14)$$

Message Framing

To perceive the uncertainty in the message framing of post p , we used the same approach as described in the ‘‘Uncertainty Detection Model’’ section to construct the uncertainty representation of the post p :

$$I_F = \text{MLP}[F(p) \boxtimes p] \quad (15)$$

Physical Environment

To measure uncertainty in the physical environment, we collected the daily number of new cases from the start of the COVID-19 outbreak and counted the number of daily news items related to the outbreak, denoted as N^{Cases} and N^{News} , respectively. Intuitively, the higher the number of new cases and news items for a day, the more sensitive the public is to the social environment and the more uncertain the environment is on that day. Thus, the uncertainty factor in the physical environment is defined as follows:

$$f_i^{\text{ph}} = \text{Norm}(\log(1 + \text{abs}(N_i^{\text{Cases}} - N_{i-1}^{\text{Cases}})) \times \log(1 + \text{abs}(N_i^{\text{News}} - N_{i-1}^{\text{News}}))) \quad (16)$$

where f_i^{ph} denotes the uncertainty factor at day i and abs is the absolute value operation. For each post, we can obtain the uncertainty factor for its corresponding date $f^{\text{ph}}(p)$.

We added the uncertainty factor of the physical environment to the representations of *macro-media environment* (I_M), *micro-communicative environment* (I_C), and *post message framing* (I_F) to get the representation of the physical environment, denoted as I_P :

$$I_P = (f^{\text{ph}} \times I_M) \boxtimes (f^{\text{ph}} \times I_C) \boxtimes (f^{\text{ph}} \times I_F) \quad (17)$$

Prediction

Prediction With EUP Alone Without Baseline Models

We concatenate the above 4 environment uncertainty features and feed the result into an MLP layer and a softmax layer for the final prediction:

$$I_{\text{EUP}} = I_M \boxtimes I_C \boxtimes I_F \boxtimes I_P \quad (18)$$



Prediction With Baseline Models

We expect that our EUP is compatible with and can empower various misinformation detection and prediction algorithms. Therefore, we used an adaptive feature selection approach based on a gate mechanism to accommodate different misinformation detectors:

$$I = g_M \cdot I_M + g_C \cdot I_C + g_F \cdot I_F + g_P \cdot I_P \quad (20)$$

where \mathbf{o} denotes the last-layer feature from the misinformation baseline algorithm. The gating vector $g_M = \text{sigmoid}(\text{Linear}(\mathbf{o} \cdot I_M))$ and g_C , g_F , and g_P are obtained in the same way. Then, we concatenate \mathbf{o} and I , and fed the result into an MLP layer and a softmax layer for the final prediction:



During training, we minimize the cross-entropy loss.

Ethical Considerations

The study is exempt from ethical review for human subject research for the following reasons. First, the study uses data from 2 publicly available Twitter data sets collected through the official application programming interface (API) of the Twitter platform for gathering tweets. The news data set was obtained from the official websites of news media. Second, the

data used in this study are anonymized and do not contain any personally identifiable information. It is also impossible to reidentify individuals from the data set. The data set is stored on a dedicated secure data server, and the analysis is conducted on the platform's designated site. This process is undertaken for research purposes and adheres to Chinese data privacy laws and regulations. Third, this study does not involve any experimental manipulation of human individuals or other ethical concerns. For instance, it does not include data on children under 18 years of age, which require legally mandated parental or guardian supervision. It also does not encompass sensitive aspects of participants' behavior or pose any physical, psychological, or economic harm or risk to the research participants.

Data Set and Experiment

Data Set

The statistics and description of our experimental data set are shown in Tables 1 and 2, respectively.

Table 1. Statistics of the data set.^{a,b}

Data set	Misinformation detection, n		Spread prediction, n		Total, n
	Real	Fake	Low	High	
Train	901	1324	1054	1171	2225
Value	312	430	360	382	742
Test	310	432	358	384	742

^aNews items in $M=58,095$. The corresponding mean and range are 988 and 10-2511, respectively.

^bTweet items in $C=321,656$. The corresponding mean and range are 793, 138-1214, respectively.

Table 2. Descriptions of the data set.

Data	Features	Size, n
Post	Content, created time, retweet count, veracity label, retweeted label	3709
News	Content, created time	58,095
Tweets	Content, created time	321,656

Post

We processed and integrated 2 existing COVID-19 data sets, FibVID [58] and CMU_MisCov19 [59], for our experiments. Both data sets have been labeled for veracity by experts, providing ground-truth labels for our experimental evaluations. For FibVID, we extracted data related to COVID-19, assigning veracity tags as 0 (COVID true) or 1 (COVID fake). We relabeled CMU_MisCov19, classifying *calling out or correction*, *true public health response*, and *true prevention* as *real* tags, and *conspiracy*, *fake cure*, *sarcasm or satire*, *false fact or prevention*, *fake treatment*, and *false public health response* as *fake* tags. Furthermore, we used the Twitter API to retrieve the number of retweets for all tweets in both data sets. Subsequently, we categorized the retweet labels as low (when the retweet count is 0) and high (when the retweet count is >0) following an analysis of the distribution of retweet numbers. The data revealed that misinformation was predominantly observed from January to July 2020, coinciding with the period of heightened uncertainty during the pandemic outbreak. Consequently, our

focus was directed solely to this specific period, resulting in the extraction of 3709 posts from January to July of 2020.

Macro-Media Environment

We gathered all the news headlines and brief descriptions from the Huffington Post, NPR, and Daily Mail from January to July 2020, as per the methodology outlined previously [12]. Notably, these 3 outlets represent the left-, center-, and right-wing perspectives, contributing to the diversity of news items for our analysis. We then used the keywords “covid,” “coronavirus,” “pneumonia,” “pandemic,” “epidemic,” “infection,” “prevalence,” and “symptom” to filter these data to ensure that the collected data were relevant to COVID-19. We ended up with 58,095 news items from January to July 2020.

Micro-Communicative Environment

We obtained the tweet IDs associated with COVID-19 from an ongoing project [60]. Given the substantial volume, we randomly sampled 1% of these IDs (amounting to approximately 205,581,778 records). Subsequently, using the Twitter API, we

retrieved the content associated with these IDs, resulting in a data set comprising 321,656 tweets spanning from January to July 2020.

Physical Environment

We compiled the daily count of new worldwide COVID-19 cases starting from January 2020, utilizing the Our World in Data database. Additionally, the daily volume of news data corresponds to the information we gathered during the same period.

Experimental Setup

Tasks

We used the proposed model for 2 tasks:

Task 1. Misinformation Detection

The objective was to analyze the text content of a tweet and ascertain whether it contained misinformation.

Textbox 1. Baseline models.

- Bidirectional long short-term memory**

Bidirectional long short-term memory (BiLSTM) [63] is a type of recurrent neural network architecture designed for sequence modeling tasks, particularly in natural language processing. It processes input sequences in both forward and backward directions simultaneously, allowing the model to capture information from both past and future contexts.

- Event adversarial neural networks**

Event adversarial neural networks (EANN_T) [64] is a model using adversarial training to eliminate event-specific features derived from a convolutional neural network for text (ie, TextCNN).

- BERT**

Bidirectional encoder representations from transformers (BERT) [52] is a pretrained language model based on deep bidirectional transformers.

- BERT-Emo**

BERT-Emo [65] is a fake news detection model that integrates multiple sentiment features into BERT.

Evaluation Metrics

For both tasks, we used accuracy and macro- F_1 -score as evaluation metrics. Additionally, in task 1, we used F_1 -scores for fake ($F_{1\text{fake}}$) and real ($F_{1\text{real}}$), while in task 2, we considered F_1 -scores for low ($F_{1\text{low}}$) and high ($F_{1\text{high}}$). Further implementation details can be found in [Multimedia Appendix 1](#).

Task 2: Spread Prediction

The objective was to evaluate the text content of a tweet to determine whether it is likely to be retweeted.

Uncertainty Features

Following Jean et al [53], we used WikiWeasel [61], a comprehensive corpus consisting of paragraphs extracted from Wikipedia, to compute the frequency of each lemma. The uncertainty score for each sentence is determined using mean pooling $F^{A,\text{Mean}}$. We leverage [62] to acquire sentence representations, relying on pretrained BERT models [52] and subsequent posttraining on news items. In the macro-media environment and the micro-communicative environment, we set $T=3$, $r=0.1$, $|C|_{\text{min}}=10$.

Baseline Models

The baseline models considered are listed in [Textbox 1](#).

Results

Overview

[Tables 3](#) and [4](#) showcase the performances of the EUP without baseline models and those of various baseline models, with and without EUP, for the misinformation detection and spread prediction tasks, respectively. The results indicate that the performances of EUP are comparable to those of state-of-the-art baseline models in both tasks. Moreover, it is noteworthy that all baseline models exhibit performance improvements when incorporating EUP for both tasks. These observations suggest the effectiveness of our proposed EUP.

Table 3. Model performance comparison on the misinformation detection task without the baseline algorithm or without the EUP^a module.^b

Model	Accuracy	Macro- F_1 -score	F_1 fake	F_1 real
EUP	<i>0.753</i>	<i>0.739</i>	<i>0.800</i>	<i>0.677</i>
BiLSTM ^c	0.733	0.729	0.783	0.683
BiLSTM + EUP	<i>0.755</i>	<i>0.743</i>	<i>0.798</i>	<i>0.688</i>
EANN _T ^d	0.745	0.730	0.795	0.664
EANN _T + EUP	<i>0.767</i>	<i>0.765</i>	<i>0.806</i>	<i>0.708</i>
BERT ^e	0.755	0.743	0.797	0.689
BERT + EUP	<i>0.771</i>	<i>0.767</i>	0.796	<i>0.738</i>
BERT-Emo	0.749	0.740	0.789	0.691
BERT-Emo + EUP	<i>0.768</i>	<i>0.763</i>	<i>0.799</i>	<i>0.726</i>

^aEUP: Environmental Uncertainty Perception.

^bThe best result in each group is in italics.

^cBiLSTM: bidirectional long short-term memory.

^dEANN_T: event adversarial neural networks.

^eBERT: bidirectional encoder representations from transformers.

Table 4. Model performance comparison on the spread prediction task without the baseline algorithm or without the EUP^a module.^b

Model	Accuracy	Macro- F_1 -score	F_1 low	F_1 high
EUP	<i>0.710</i>	<i>0.710</i>	<i>0.719</i>	<i>0.701</i>
BiLSTM ^c	0.707	0.705	0.684	0.726
BiLSTM + EUP	<i>0.734</i>	<i>0.733</i>	<i>0.738</i>	<i>0.729</i>
EANN _T ^d	0.717	0.716	0.734	0.698
EANN _T + EUP	<i>0.726</i>	<i>0.726</i>	<i>0.736</i>	<i>0.716</i>
BERT ^e	0.728	0.728	0.728	0.728
BERT + EUP	<i>0.743</i>	<i>0.743</i>	<i>0.752</i>	<i>0.734</i>
BERT-Emo	0.733	0.733	0.730	0.737
BERT-Emo + EUP	<i>0.741</i>	<i>0.741</i>	<i>0.733</i>	<i>0.749</i>

^aEUP: Environmental Uncertainty Perception.

^bThe best result in each group is in italics.

^cBiLSTM: bidirectional long short-term memory.

^dEANN_T: event adversarial neural networks.

^eBERT: bidirectional encoder representations from transformers.

Ablation Study

We systematically eliminated individual components, namely, macro-media environment, micro-communicative environment, message framing, and physical environment, and assessed the modeling performances on the data set. [Tables 5](#) and [6](#) illustrate

that, under all experimental conditions, performance degrades when any of these components are removed. These results underscore the effectiveness of all 4 uncertainty features of the information environment for both misinformation detection and spread prediction.

Table 5. Ablation study on the misinformation detection task.^a

Model	Accuracy	Macro- F_1 -score	F_1 fake	F_1 real
EUP^b	<i>0.753</i>	<i>0.739</i>	<i>0.800</i>	<i>0.677</i>
Without I_M	0.748	0.738	0.790	0.687
Without I_C	0.745	0.720	0.803	0.637
Without I_F	0.739	0.734	0.778	0.673
Without I_P	0.747	0.730	0.797	0.663
BiLSTM^c + EUP	<i>0.755</i>	<i>0.743</i>	<i>0.798</i>	<i>0.688</i>
Without I_M	0.745	0.741	0.793	0.669
Without I_C	0.741	0.728	0.788	0.668
Without I_F	0.747	0.735	0.791	0.678
Without I_P	0.746	0.742	0.796	0.665
BERT^d + EUP	<i>0.771</i>	<i>0.767</i>	<i>0.796</i>	<i>0.738</i>
Without I_M	0.762	0.754	0.801	0.707
Without I_C	0.764	0.761	0.807	0.696
Without I_F	0.761	0.752	0.800	0.705
Without I_P	0.758	0.751	0.795	0.707

^aThe best result in each group is in italics.

^bEUP: Environmental Uncertainty Perception.

^cBiLSTM: bidirectional long short-term memory.

^dBERT: bidirectional encoder representations from transformers.

Table 6. Ablation study on the spread prediction task.^a

Model	Accuracy	Macro- F_1 -score	F_1 low	F_1 high
EUP^b	<i>0.710</i>	<i>0.710</i>	0.719	<i>0.701</i>
Without I_M	0.697	0.696	0.715	0.676
Without I_C	0.695	0.694	0.712	0.677
Without I_F	0.702	0.702	0.714	0.689
Without I_P	0.708	0.707	<i>0.721</i>	0.692
BiLSTM^c + EUP	<i>0.734</i>	<i>0.733</i>	0.738	<i>0.729</i>
Without I_M	0.724	0.723	0.735	0.711
Without I_C	0.721	0.721	0.716	0.726
Without I_F	0.717	0.716	0.731	0.702
Without I_P	0.726	0.723	<i>0.753</i>	0.693
BERT^d + EUP	<i>0.743</i>	<i>0.743</i>	0.752	<i>0.734</i>
Without I_M	0.741	0.739	0.764	0.713
Without I_C	0.741	0.738	<i>0.766</i>	0.711
Without I_F	0.736	0.735	0.753	0.716
Without I_P	0.740	0.738	0.759	0.717

^aThe best result in each group is in italics.

^bEUP: Environmental Uncertainty Perception.

^cBiLSTM: bidirectional long short-term memory.

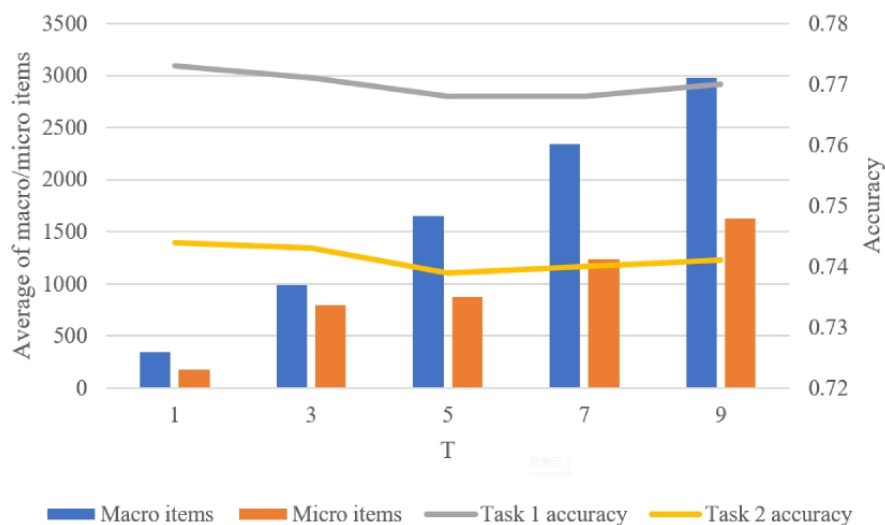
^dBERT: bidirectional encoder representations from transformers.

The Effect of the Day Parameter T

To explore the impact of the day parameter (T) on the results during the construction of the macro-media environment and the micro-communicative environment, we experimented with different values of T . Specifically, we sequentially set $T=1, 3,$

5, 7, and 9 for the BERT + EUP model, and the experimental results are depicted in [Figure 2](#). Despite the fact that increasing T results in larger macro-media and micro-communicative environments, the optimal performance was achieved when $T=1$.

Figure 2. The effect of the day parameter T . Lines show the accuracies of both tasks and bars show the average number of news and tweet items in the environments.

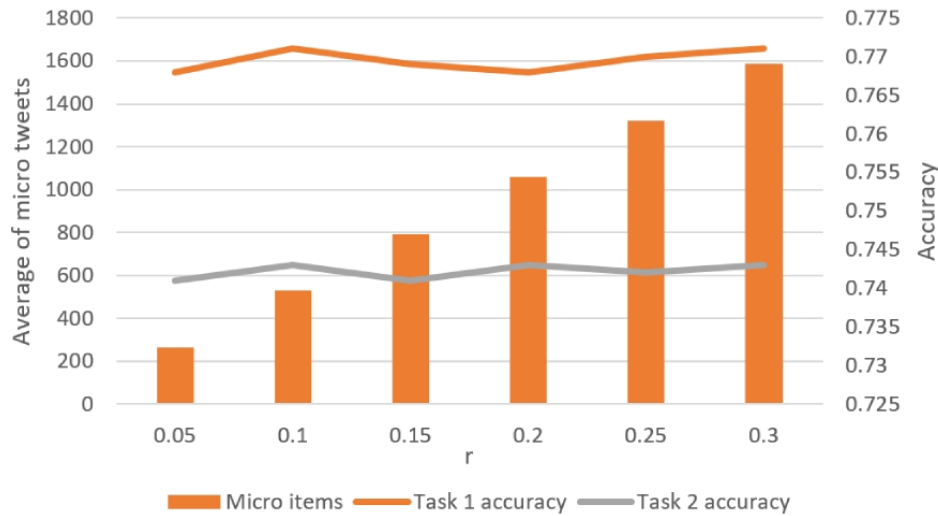


The Effect of the Rate Parameter r

We maintained the setting $T=3$ and systematically varied r , using values of 0.05, 0.1, 0.15, 0.2, 0.25, and 0.3 on the BERT

+ EUP model to examine the impact of r on the experimental results, as illustrated in Figure 3. The accuracy performance exhibited fluctuations with varying values of r . Notably, the highest accuracy for both tasks was observed when $r=0.1$.

Figure 3. The effect of the rate parameter r . Lines show the accuracies of both tasks and bars show the average number of tweet items in the environment.

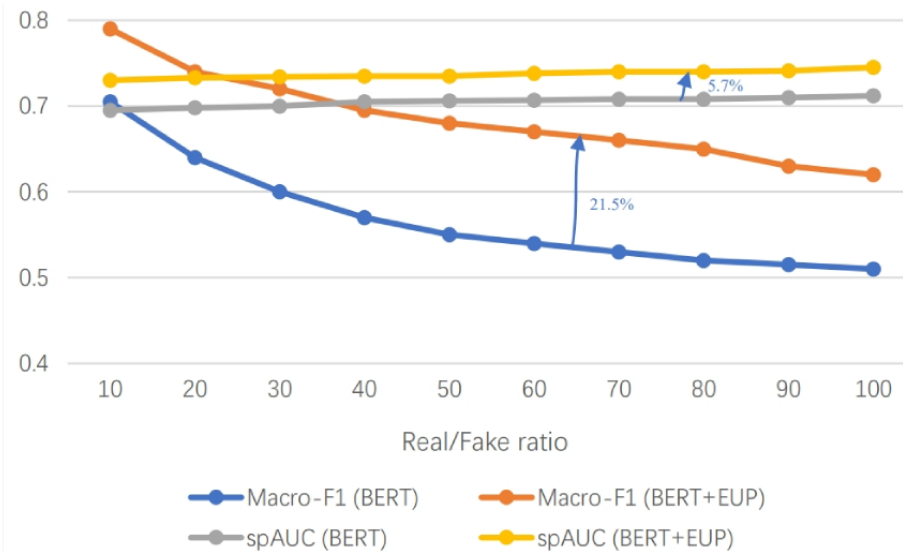


Evaluation on Imbalanced Data

In real-world scenarios, the distribution of real and fake information often exhibits significant imbalance. To evaluate the efficacy of our proposed EUP framework on unbalanced data sets, we conducted tests on data sets with varying ratios of real to fake data, ranging from 10:1 to 100:1. We measured and

reported macro- F_1 -scores and standardized partial area under the curve (AUC) with a false-positive rate of at most 0.1 (ie, $spAUCFPR \leq 0.1$ [66]) to assess the effectiveness of our EUP framework in handling nonbalanced data sets. As depicted in Figure 4, EUP yields relative improvements of 21.5% and 5.7% in macro- F_1 -score and $spAUCFPR \leq 0.1$, demonstrating its effectiveness on unbalanced data sets.

Figure 4. Performance of macroF1 and spAUC values across datasets with varying ratios.



Discussion

Principal Findings

First, this study enhances scholars' comprehension of the misinformation detection and spread prediction problem by highlighting the significance of uncertainty in information environments. Notably, this research contributes to the literature by recognizing uncertainty features in the information environments of misinformation as a pivotal factor for

improving detection and prediction algorithms during a pandemic. Our findings underscore that the EUP alone is sufficient for both tasks and has the potential to enhance the capabilities of state-of-the-art algorithms. In contrast to prior misinformation research that primarily concentrates on post content (such as post theme, sentiments, and linguistic characteristics, as seen in [6,11,29]) and network connections (eg, number of followers [25]) on social media, this study advances scholars' understanding of the misinformation problem by emphasizing the importance of uncertainty in information

environments. Recognizing and incorporating uncertainty as a fundamental concept in misinformation detection and spread prediction during crises hold theoretical significance. This is particularly relevant as a crisis is characterized by its unpredictable, unexpected, and nonroutine nature, inherently giving rise to uncertainty [38,67]. This uncertainty has been theorized to compel individuals to seek information as a coping mechanism for dealing with the anxiety and pressure generated by uncertainty. This process allows people to diminish uncertainty, restore a sense of normalcy, and alleviate anxiety [14,68]. Regrettably, this coping mechanism can inadvertently fuel the proliferation and dissemination of misinformation, particularly when there is a lack of timely and accurate information, contributing to the concurrent occurrence of an infodemic [6,11,50]. The current research seeks to advance the literature by establishing the legitimacy of uncertainty in the information environments of misinformation as a central indicator for the detection and prediction of misinformation during public health crises.

Second, this study delves into the intricacies of uncertain information environments for misinformation across 4 distinct scales, namely, the physical environment, macro-media environment, micro-communicative environment, and message framing. Our findings demonstrate the effectiveness of all 4 uncertainty features in misinformation detection and spread prediction. In contrast to prior misinformation literature during the COVID-19 pandemic, which often overlooked the role of the information environment in increasing the likelihood of misinformation dissemination, our research emphasizes the importance of considering uncertainty beyond the content of misinformation itself, such as ambiguous wording [29,50]. Our study broadens the concept of linguistic uncertainty in misinformation message framing to encompass a more comprehensive uncertainty across various information environments. We define uncertainty in information environments using a multiscale approach that highlights the significance of the interaction between the physical environment and macro-/micro-media environments. This approach diverges from focusing on a single dimension, such as ambiguities about official guidelines and news reports [18], or the misinformation framing strategy on social media [29].

Third, our findings indicate that uncertainties in information environments play a crucial role as motivators for the emergence and spread of misinformation. While previous studies have provided preliminary evidence suggesting that uncertainty stemming from government policies and news media could coincide with the occurrence of related misinformation during the COVID-19 pandemic, often relying on descriptive big data analyses [3,32], our study contributes stronger empirical evidence. We leverage machine learning techniques to demonstrate that uncertainty arising from the crisis and crisis communication through media can indeed incentivize individuals to generate and disseminate misinformation. Significantly, our findings revealed that the algorithm achieved its best performance for both detection and spread prediction tasks when incorporating items from the information environments published 1 day before the post ($T=1$). This discovery emphasizes the acute impact of uncertainty in the

information environment on the emergence and spread of misinformation, underscoring the importance of timely uncertainty reduction in crisis communication. Furthermore, the algorithm attained the highest accuracies when it included items highly relevant to the post but with an appropriate size ($r=0.1$). This rationale is reasonable, as a too-small r may fail to encompass enough misinformation-related items, while a larger r might include a significant amount of irrelevant information. The evidence theoretically establishes a connection between crisis communication research and misinformation research, reinforcing the notion that crisis communication and misinformation containment are 2 intertwined aspects of crisis management [3].

This study offers significant practical implications for misinformation detection and spread prediction. First, unlike previous studies that separately investigated computational frameworks for these tasks [24,29], this study introduces a unified uncertainty-based framework capable of addressing both tasks simultaneously. Second, our framework operates instantaneously, as it only requires easily accessible data such as posts, mainstream news, and relevant social media discussions published a few days prior. Moreover, the uncertainty detection algorithm has been trained using external data, rendering our algorithm easy to implement and capable of providing timely detection and prediction for streaming textual data. Third, this study affirms the effectiveness of uncertainty in various information environments for detecting and predicting misinformation on social media. Hence, the 4 proposed uncertainty components in information environments could be leveraged by social media platforms to improve the accuracy of misinformation detection and spread prediction, thereby safeguarding individuals from harm caused by infodemic. The benefits offered by our algorithm may serve as an impetus for integrating uncertainty components into practical systems.

Limitations and Future Work

This study is the first to incorporate the uncertainty present in the information environment of a post for both misinformation detection and spread prediction. However, it has some limitations. First, our framework concentrated solely on text-only detection and prediction. Future work should extend the framework to incorporate multimodal and social graph-based detection. Second, we used an uncertainty detection algorithm developed from a generic corpus sourced from Wikipedia. Nevertheless, past research has indicated that expressions of uncertainty may vary slightly across domains [53]. In other words, uncertainty expressions in the context of the COVID-19 pandemic may differ from those in general situations. Therefore, future work should aim to enhance our uncertainty measure by utilizing a corpus specifically designed for uncertainty detection in the discourse related to COVID-19.

Conclusions

We introduced an EUP framework for both misinformation detection and spread prediction. Our framework delves into uncertainty within information environments across 4 scales: the physical environment, macro-media environment, micro-communicative environment, and message framing. The experiments demonstrated the effectiveness of our proposed

uncertainty components in enhancing the performance of existing models. There are several directions for further investigation and extension of this work. First, we can explore the impact of different news and social media environments (eg, biased vs neutral; left wing vs right wing) on the emergence and spread of misinformation. Second, extending our algorithms to include multimodal misinformation detection could be

beneficial, as misinformation increasingly incorporates images and videos. Third, investigating the interaction between misinformation detection and spread prediction using a multitask, transfer-learning model is a promising avenue, given the shared uncertainty framework identified in this study for both tasks.

Acknowledgments

This study was supported by Open Funding Project of the State Key Laboratory of Communication Content Cognition (grant number 20G01).

Data Availability

The data sets generated and analyzed during this study are available from the corresponding author on reasonable request.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Uncertainty features.

[[DOCX File , 18 KB - ai_v3i1e47240_app1.docx](#)]

References

1. Thomas Z. WHO says fake coronavirus claims causing "infodemic". BBC. 2020. URL: <https://www.bbc.com/news/technology-51497800> [accessed 2022-09-08]
2. Bode L, Vraga EK. See something, say something: correction of global health misinformation on social media. *Health Commun* 2018 Sep 16;33(9):1131-1140. [doi: [10.1080/10410236.2017.1331312](https://doi.org/10.1080/10410236.2017.1331312)] [Medline: [28622038](https://pubmed.ncbi.nlm.nih.gov/28622038/)]
3. Lu J. Themes and evolution of misinformation during the early phases of the COVID-19 outbreak in China—an application of the crisis and emergency risk communication model. *Front Commun* 2020 Aug 14;5:57. [doi: [10.3389/fcomm.2020.00057](https://doi.org/10.3389/fcomm.2020.00057)]
4. Swire-Thompson B, Lazer D. Public health and online misinformation: challenges and recommendations. *Annu Rev Public Health* 2020 Apr 02;41(1):433-451 [FREE Full text] [doi: [10.1146/annurev-publhealth-040119-094127](https://doi.org/10.1146/annurev-publhealth-040119-094127)] [Medline: [31874069](https://pubmed.ncbi.nlm.nih.gov/31874069/)]
5. Jiang G, Liu S, Zhao Y, Sun Y, Zhang M. Fake news detection via knowledgeable prompt learning. *Information Processing & Management* 2022 Sep;59(5):103029. [doi: [10.1016/j.ipm.2022.103029](https://doi.org/10.1016/j.ipm.2022.103029)]
6. Kumari R, Ashok N, Ghosal T, Ekbal A. Misinformation detection using multitask learning with mutual learning for novelty detection and emotion recognition. *Information Processing & Management* 2021 Sep;58(5):102631. [doi: [10.1016/j.ipm.2021.102631](https://doi.org/10.1016/j.ipm.2021.102631)]
7. Babic K. Prediction of COVID-19 Related Information Spreading on Twitter. New York, NY: IEEE; 2021 Sep 27 Presented at: 2021 44th International Convention on Information, Communication and Electronic Technology (MIPRO); May 24-28, 2021; Opatija, Croatia p. 395-399. [doi: [10.23919/MIPRO52101.2021.9596693](https://doi.org/10.23919/MIPRO52101.2021.9596693)]
8. Ghina Khoerunnisa, Jondri, Widi Astuti. Prediction of retweets based on user, content, and time features using EUSBoost. *J RESTI (Rekayasa Sist Teknol Inf)* 2022 Jun 30;6(3):442-447. [doi: [10.29207/resti.v6i3.4125](https://doi.org/10.29207/resti.v6i3.4125)]
9. Islam MR, Liu S, Wang X, Xu G. Deep learning for misinformation detection on online social networks: a survey and new perspectives. *Soc Netw Anal Min* 2020;10(1):82 [FREE Full text] [doi: [10.1007/s13278-020-00696-x](https://doi.org/10.1007/s13278-020-00696-x)] [Medline: [33014173](https://pubmed.ncbi.nlm.nih.gov/33014173/)]
10. Shu K, Sliva A, Wang S, Tang J, Liu H. Fake news detection on social media. *SIGKDD Explor Newsl* 2017 Sep;19(1):22-36. [doi: [10.1145/3137597.3137600](https://doi.org/10.1145/3137597.3137600)]
11. Su Q, Wan M, Liu X, Huang C. Motivations, methods and metrics of misinformation detection: an NLP perspective. *NLPRE* 2020;1(1-2):1. [doi: [10.2991/nlpr.d.200522.001](https://doi.org/10.2991/nlpr.d.200522.001)]
12. Sheng Q, Cao J, Zhang X, Li R, Wang D, Zhu Y. Zoom out and observe: news environment perception for fake news detection. New York, NY: Association for Computational Linguistics; 2022 Presented at: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers); May 22-27, 2022; Dublin, Ireland p. 4543-4556. [doi: [10.18653/v1/2022.acl-long.311](https://doi.org/10.18653/v1/2022.acl-long.311)]
13. Rosnow R. Rumor as communication: a contextualist approach. *Journal of Communication* 1988;38(1):12-28. [doi: [10.1111/j.1460-2466.1988.tb02033.x](https://doi.org/10.1111/j.1460-2466.1988.tb02033.x)]
14. Bradac JJ. Theory comparison: uncertainty reduction, problematic integration, uncertainty management, and other curious constructs. *Journal of Communication* 2001;51(3):456-476. [doi: [10.1111/j.1460-2466.2001.tb02891.x](https://doi.org/10.1111/j.1460-2466.2001.tb02891.x)]

15. Oh O, Agrawal M, Rao HR. Community intelligence and social media services: a rumor theoretic analysis of tweets during social crises. *MISQ* 2013 Feb 2;37(2):407-426. [doi: [10.25300/misq/2013/37.2.05](https://doi.org/10.25300/misq/2013/37.2.05)]
16. Tandoc EC, Lee JCB. When viruses and misinformation spread: how young Singaporeans navigated uncertainty in the early stages of the COVID-19 outbreak. *New Media & Society* 2020 Oct 25;24(3):778-796. [doi: [10.1177/1461444820968212](https://doi.org/10.1177/1461444820968212)]
17. Capurro G, Jardine CG, Tustin J, Driedger M. Communicating scientific uncertainty in a rapidly evolving situation: a framing analysis of Canadian coverage in early days of COVID-19. *BMC Public Health* 2021 Nov 29;21(1):2181-2114 [FREE Full text] [doi: [10.1186/s12889-021-12246-x](https://doi.org/10.1186/s12889-021-12246-x)] [Medline: [34844582](https://pubmed.ncbi.nlm.nih.gov/34844582/)]
18. Zhang YSD, Young Leslie H, Sharafaddin-Zadeh Y, Noels K, Lou NM. Public health messages about face masks early in the COVID-19 pandemic: perceptions of and impacts on Canadians. *J Community Health* 2021 Oct 20;46(5):903-912 [FREE Full text] [doi: [10.1007/s10900-021-00971-8](https://doi.org/10.1007/s10900-021-00971-8)] [Medline: [33611755](https://pubmed.ncbi.nlm.nih.gov/33611755/)]
19. Dietrich AM, Kuester K, Müller GJ, Schoenle R. News and uncertainty about COVID-19: survey evidence and short-run economic impact. *J Monet Econ* 2022 Jul;129:S35-S51 [FREE Full text] [doi: [10.1016/j.jmoneco.2022.02.004](https://doi.org/10.1016/j.jmoneco.2022.02.004)] [Medline: [35165494](https://pubmed.ncbi.nlm.nih.gov/35165494/)]
20. Cao J, Qi P, Sheng Q, Yang T, Guo J, Li J. Exploring the role of visual content in fake news detection. In: Shu K, Wang S, Lee D, Liu H, editors. *Disinformation, Misinformation, and Fake News in Social Media*. Cham, Switzerland: Springer; Jun 18, 2020:141-161.
21. Qi P, Cao J, Li X, Liu H, Sheng Q, Mi X, et al. Improving fake news detection by using an entity-enhanced framework to fuse diverse multimodal clues. In: *MM '21: Proceedings of the 29th ACM International Conference on Multimedia*. New York, NY: Association for Computing Machinery; 2021 Oct Presented at: The 29th ACM International Conference on Multimedia (MM '21); October 17, 2021; Chengdu, China p. 1212-1220. [doi: [10.1145/3474085.3481548](https://doi.org/10.1145/3474085.3481548)]
22. Liu C, Wu X, Yu M, Li G, Jiang J, Huang W, et al. A two-stage model based on BERT for short fake news detection. Cham, Switzerland: Springer; 2019 Aug 22 Presented at: International Conference on Knowledge Science, Engineering and Management (KSEM 2019); August 28-30, 2019; Athens, Greece p. 172-183. [doi: [10.1007/978-3-030-29563-9_17](https://doi.org/10.1007/978-3-030-29563-9_17)]
23. Vo N, Lee K. Hierarchical multi-head attentive network for evidence-aware fake news detection. In: *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics*. New York, NY: Association for Computational Linguistics; 2021 Apr Presented at: The 16th Conference of the European Chapter of the Association for Computational Linguistics; April 1, 2021; Online p. 965-975. [doi: [10.18653/v1/2021.eacl-main.83](https://doi.org/10.18653/v1/2021.eacl-main.83)]
24. Silva A, Han Y, Luo L, Karunasekera S, Leckie C. Propagation2Vec: embedding partial propagation networks for explainable fake news early detection. *Information Processing & Management* 2021 Sep;58(5):102618. [doi: [10.1016/j.ipm.2021.102618](https://doi.org/10.1016/j.ipm.2021.102618)]
25. Zhao Y, Da J, Yan J. Detecting health misinformation in online health communities: incorporating behavioral features into machine learning based approaches. *Information Processing & Management* 2021 Jan;58(1):102390. [doi: [10.1016/j.ipm.2020.102390](https://doi.org/10.1016/j.ipm.2020.102390)]
26. Shaden S, Nikolay B, Giovanni DSM, Preslav N. That is a known lie: detecting previously fact-checked claims. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. New York, NY: Association for Computational Linguistics; 2020 Presented at: The 58th Annual Meeting of the Association for Computational Linguistics; July 5-10, 2020; Online p. 3607-3618 URL: <https://aclanthology.org/2020.acl-main.332.pdf> [doi: [10.18653/v1/2020.acl-main.332](https://doi.org/10.18653/v1/2020.acl-main.332)]
27. Cui L, Seo H, Tabar M, Ma F, Wang S, Lee D. DETERRENT: knowledge guided graph attention network for detecting healthcare misinformation. In: *KDD '20: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. New York, NY: Association for Computing Machinery; 2020 Aug 20 Presented at: KDD '20: The 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining; July 6-10, 2020; Virtual Event. [doi: [10.1145/3394486.3403092](https://doi.org/10.1145/3394486.3403092)]
28. Kumar KPK, Geethakumari G. Detecting misinformation in online social networks using cognitive psychology. *Hum Cent Comput Inf Sci* 2014 Sep 24;4(1):1-22. [doi: [10.1186/s13673-014-0014-x](https://doi.org/10.1186/s13673-014-0014-x)]
29. Zhou C, Li K, Lu Y. Linguistic characteristics and the dissemination of misinformation in social media: the moderating effect of information richness. *Information Processing & Management* 2021 Nov;58(6):102679. [doi: [10.1016/j.ipm.2021.102679](https://doi.org/10.1016/j.ipm.2021.102679)]
30. Keller AC, Ansell CK, Reingold AL, Bourrier M, Hunter MD, Burrowes S, et al. Improving pandemic response: a sensemaking perspective on the spring 2009 H1N1 pandemic. *Risk Hazard & Crisis Pub Pol* 2012 Aug 10;3(2):1-37. [doi: [10.1515/1944-4079.1101](https://doi.org/10.1515/1944-4079.1101)]
31. Genuis SK. Constructing “sense” from evolving health information: a qualitative investigation of information seeking and sense making across sources. *J Am Soc Inf Sci Tec* 2012 Jun 29;63(8):1553-1566. [doi: [10.1002/asi.22691](https://doi.org/10.1002/asi.22691)]
32. Lu J, Zhang M, Zheng Y, Li Q. Communication of uncertainty about preliminary evidence and the spread of its inferred misinformation during the COVID-19 pandemic—a Weibo case study. *Int J Environ Res Public Health* 2021 Nov 13;18(22):11933 [FREE Full text] [doi: [10.3390/ijerph182211933](https://doi.org/10.3390/ijerph182211933)] [Medline: [34831688](https://pubmed.ncbi.nlm.nih.gov/34831688/)]
33. Heverin T, Zach L. Use of microblogging for collective sense - making during violent crises: a study of three campus shootings. *J Am Soc Inf Sci* 2011 Oct 24;63(1):34-47. [doi: [10.1002/asi.21685](https://doi.org/10.1002/asi.21685)]

34. Kim HK, Ahn J, Atkinson L, Kahlor LA. Effects of COVID-19 misinformation on information seeking, avoidance, and processing: a multicountry comparative study. *Science Communication* 2020 Sep 13;42(5):586-615. [doi: [10.1177/1075547020959670](https://doi.org/10.1177/1075547020959670)]
35. Vos SC, Buckner MM. Social media messages in an emerging health crisis: tweeting bird flu. *J Health Commun* 2016 Dec 31;21(3):301-308. [doi: [10.1080/10810730.2015.1064495](https://doi.org/10.1080/10810730.2015.1064495)] [Medline: [26192209](https://pubmed.ncbi.nlm.nih.gov/26192209/)]
36. Wood S, Michaelides G, Daniels K, Niven K. Uncertainty and well-being amongst homeworkers in the COVID-19 pandemic: a longitudinal study of university staff. *Int J Environ Res Public Health* 2022 Aug 22;19(16):10435 [FREE Full text] [doi: [10.3390/ijerph191610435](https://doi.org/10.3390/ijerph191610435)] [Medline: [36012069](https://pubmed.ncbi.nlm.nih.gov/36012069/)]
37. Longstaff PH, Yang S. Communication management and trust: their role in building resilience to "surprises" such as natural disasters, pandemic flu, and terrorism. *E&S* 2008;13(1):3-3. [doi: [10.5751/es-02232-130103](https://doi.org/10.5751/es-02232-130103)]
38. Reynolds B, W Seeger M. Crisis and emergency risk communication as an integrative model. *J Health Commun* 2005 Feb 23;10(1):43-55. [doi: [10.1080/10810730590904571](https://doi.org/10.1080/10810730590904571)] [Medline: [15764443](https://pubmed.ncbi.nlm.nih.gov/15764443/)]
39. Fink S. *Crisis Management: Planning for the Inevitable*. New York, NY: AMACOM; 1986.
40. Lwin M, Lu J, Sheldenkar A, Schulz P. Strategic uses of Facebook in zika outbreak communication: implications for the crisis and emergency risk communication model. *Int J Environ Res Public Health* 2018 Sep 10;15(9):1974 [FREE Full text] [doi: [10.3390/ijerph15091974](https://doi.org/10.3390/ijerph15091974)] [Medline: [30201929](https://pubmed.ncbi.nlm.nih.gov/30201929/)]
41. Lwin MO, Lu J, Sheldenkar A, Cayabyab YM, Yee AZH, Smith HE. Temporal and textual analysis of social media on collective discourses during the Zika virus pandemic. *BMC Public Health* 2020 May 29;20(1):804-809 [FREE Full text] [doi: [10.1186/s12889-020-08923-y](https://doi.org/10.1186/s12889-020-08923-y)] [Medline: [32471495](https://pubmed.ncbi.nlm.nih.gov/32471495/)]
42. Al-Zaman MS. Prevalence and source analysis of COVID-19 misinformation in 138 countries. *IFLA Journal* 2021 Aug 27;48(1):189-204. [doi: [10.1177/03400352211041135](https://doi.org/10.1177/03400352211041135)]
43. Rajan D, Koch K, Rohrer K, Bajnoczki C, Socha A, Voss M, et al. Governance of the Covid-19 response: a call for more inclusive and transparent decision-making. *BMJ Glob Health* 2020 May 05;5(5):e002655 [FREE Full text] [doi: [10.1136/bmjgh-2020-002655](https://doi.org/10.1136/bmjgh-2020-002655)] [Medline: [32371570](https://pubmed.ncbi.nlm.nih.gov/32371570/)]
44. Ahmed MS, Aurpa TT, Anwar MM. Detecting sentiment dynamics and clusters of Twitter users for trending topics in COVID-19 pandemic. *PLoS One* 2021 Aug 9;16(8):e0253300 [FREE Full text] [doi: [10.1371/journal.pone.0253300](https://doi.org/10.1371/journal.pone.0253300)] [Medline: [34370730](https://pubmed.ncbi.nlm.nih.gov/34370730/)]
45. Zhao Y, Cheng S, Yu X, Xu H. Chinese public's attention to the COVID-19 epidemic on social media: observational descriptive study. *J Med Internet Res* 2020 May 04;22(5):e18825 [FREE Full text] [doi: [10.2196/18825](https://doi.org/10.2196/18825)] [Medline: [32314976](https://pubmed.ncbi.nlm.nih.gov/32314976/)]
46. Featherstone JD, Zhang J. Feeling angry: the effects of vaccine misinformation and refutational messages on negative emotions and vaccination attitude. *J Health Commun* 2020 Sep 01;25(9):692-702. [doi: [10.1080/10810730.2020.1838671](https://doi.org/10.1080/10810730.2020.1838671)] [Medline: [33103600](https://pubmed.ncbi.nlm.nih.gov/33103600/)]
47. Higgins-Dunn N. A dog in Hong Kong tests positive for the coronavirus, WHO officials confirm. *CNBC*. 2022. URL: <https://www.cnn.com/2020/02/28/a-dog-in-hong-kong-tests-positive-for-the-coronavirus-who-confirms.html> [accessed 2020-02-28]
48. van der Bles AM, van der Linden S, Freeman ALJ, Spiegelhalter DJ. The effects of communicating uncertainty on public trust in facts and numbers. *Proc Natl Acad Sci U S A* 2020 Apr 07;117(14):7672-7683 [FREE Full text] [doi: [10.1073/pnas.1913678117](https://doi.org/10.1073/pnas.1913678117)] [Medline: [32205438](https://pubmed.ncbi.nlm.nih.gov/32205438/)]
49. Zhang X, Ghorbani AA. An overview of online fake news: characterization, detection, and discussion. *Information Processing & Management* 2020 Mar;57(2):102025. [doi: [10.1016/j.ipm.2019.03.004](https://doi.org/10.1016/j.ipm.2019.03.004)]
50. Zhou C, Xiu H, Wang Y, Yu X. Characterizing the dissemination of misinformation on social media in health emergencies: an empirical study based on COVID-19. *Inf Process Manag* 2021 Jul;58(4):102554 [FREE Full text] [doi: [10.1016/j.ipm.2021.102554](https://doi.org/10.1016/j.ipm.2021.102554)] [Medline: [36570740](https://pubmed.ncbi.nlm.nih.gov/36570740/)]
51. Liu Y, Ren C, Shi D, Li K, Zhang X. Evaluating the social value of online health information for third-party patients: is uncertainty always bad? *Information Processing & Management* 2020 Sep;57(5):102259. [doi: [10.1016/j.ipm.2020.102259](https://doi.org/10.1016/j.ipm.2020.102259)]
52. Devlin J, Chang M, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. In: *Proceedings of NAACL-HLT 2019*. New York, NY: Association for Computational Linguistics; 2019 Presented at: 2019 Conference of the North American Chapter of the Association for Computational Linguistics; June 2-7, 2019; Minneapolis, MN p. 4171-4186 URL: <https://aclanthology.org/N19-1423.pdf> [doi: [10.18653/v1/n18-2](https://doi.org/10.18653/v1/n18-2)]
53. Jean PA, Harispe S, Ranwez S, Bellot P, Montmain J. Uncertainty detection in natural language: a probabilistic model. In: *WIMS '16: Proceedings of the 6th International Conference on Web Intelligence, Mining and Semantics*. New York, NY: Association for Computing Machinery; 2016 Jun 13 Presented at: WIMS '16: International Conference on Web Intelligence, Mining and Semantics; June 13-15, 2016; Nîmes, France p. 1-10. [doi: [10.1145/2912845.2912873](https://doi.org/10.1145/2912845.2912873)]
54. Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library. *Advances in Neural Information Processing Systems* 2019:8037.
55. Loshchilov I. Decoupled Weight Decay Regularization. 2017 Nov 14 Presented at: International Conference on Learning Representations; 2018; online.

56. Pennington J, Socher R, Manning C. GloVe: Global Vectors for Word Representation. 2014 Presented at: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP); 2014; Doha, Qatar. [doi: [10.3115/v1/d14-1162](https://doi.org/10.3115/v1/d14-1162)]
57. Hartigan JA, Wong MA. Algorithm AS 136: a K-means clustering algorithm. *Applied Statistics* 1979;28(1):100. [doi: [10.2307/2346830](https://doi.org/10.2307/2346830)]
58. Kim J, Aum J, Lee S, Jang Y, Park E, Choi D. FibVID: comprehensive fake news diffusion dataset during the COVID-19 period. *Telemat Inform* 2021 Nov;64:101688 [FREE Full text] [doi: [10.1016/j.tele.2021.101688](https://doi.org/10.1016/j.tele.2021.101688)] [Medline: [36567815](https://pubmed.ncbi.nlm.nih.gov/36567815/)]
59. Memon S, Carley K. Characterizing COVID-19 misinformation communities using a novel twitter dataset. In: Proceedings of the CIKM 2020 Workshops. 2020 Aug 3 Presented at: CIKM 2020 Workshops; October 19-20, 2020; Galway, Ireland p. 1-9 URL: <https://ceur-ws.org/Vol-2699/paper40.pdf>
60. Chen E, Lerman K, Ferrara E. Tracking social media discourse about the COVID-19 pandemic: development of a public coronavirus twitter data set. *JMIR Public Health Surveill* 2020 May 29;6(2):e19273 [FREE Full text] [doi: [10.2196/19273](https://doi.org/10.2196/19273)] [Medline: [32427106](https://pubmed.ncbi.nlm.nih.gov/32427106/)]
61. Farkas R, Vincze V, Szarvas G, Móra G, Csirik J. Learning to detect hedges and their scope in natural language text. In: CoNLL '10: Shared Task: Proceedings of the Fourteenth Conference on Computational Natural Language Learnin. New York, NY: Association for Computational Linguistics; 2010 Presented at: CoNLL '10: Shared Task: The Fourteenth Conference on Computational Natural Language Learnin; July 15-16, 2010; Uppsala, Sweden p. 1-12. [doi: [10.3115/1596409](https://doi.org/10.3115/1596409)]
62. Gao T, Yao X, Chen D. SimCSE: simple contrastive learning of sentence embeddings. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. 2021 Presented at: The 2021 Conference on Empirical Methods in Natural Language Processing; November 7-11, 2021; Online and Punta Cana, Dominican Republic p. 6894-6910. [doi: [10.18653/v1/2021.emnlp-main.552](https://doi.org/10.18653/v1/2021.emnlp-main.552)]
63. Graves A, Schmidhuber J. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural Netw* 2005 Jul;18(5-6):602-610. [doi: [10.1016/j.neunet.2005.06.042](https://doi.org/10.1016/j.neunet.2005.06.042)] [Medline: [16112549](https://pubmed.ncbi.nlm.nih.gov/16112549/)]
64. Wang Y, Jin Z, Yuan Y, Xun G, Jha K, Su L, et al. EANN: event adversarial neural networks for multi-modal fake news detection. In: KDD '18: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. New York, NY: Association for Computing Machinery; 2018 Presented at: KDD '18: The 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; August 19-23, 2018; London, UK p. 849-857. [doi: [10.1145/3219819.3219903](https://doi.org/10.1145/3219819.3219903)]
65. Zhang X, Cao L, Li X, Sheng Q, Zhong L, Shu K. Mining Dual Emotion for Fake News Detection. 2021 Presented at: WWW '21: Proceedings of the Web Conference 2021; 2021; New York, NY, United States p. 3465-3476. [doi: [10.1145/3442381.3450004](https://doi.org/10.1145/3442381.3450004)]
66. McClish DK. Analyzing a portion of the ROC curve. *Med Decis Making* 1989;9(3):190-195. [doi: [10.1177/0272989X8900900307](https://doi.org/10.1177/0272989X8900900307)] [Medline: [2668680](https://pubmed.ncbi.nlm.nih.gov/2668680/)]
67. Xiao Y, Cauberghe V, Hudders L. Moving forward: the effectiveness of online apologies framed with hope on negative behavioural intentions in crises. *Journal of Business Research* 2020 Mar;109:621-636. [doi: [10.1016/j.jbusres.2019.06.034](https://doi.org/10.1016/j.jbusres.2019.06.034)]
68. Brashers D. Communication and uncertainty management. *Journal of Communication* 2001;51(3):477-497. [doi: [10.1111/j.1460-2466.2001.tb02892.x](https://doi.org/10.1111/j.1460-2466.2001.tb02892.x)]

Abbreviations

- API:** application programming interface
- AUC:** area under the curve
- BERT:** bidirectional encoder representations from transformers
- BiLSTM:** bidirectional long short-term memory
- EANNT:** event adversarial neural networks
- EUP:** Environmental Uncertainty Perception
- MLP:** multilayer perceptron
- spAUCFPR:** standardized partial area under the curve with a false-positive rate
- TextCNN:** convolutional neural network for text

Edited by K El Emam, B Malin; submitted 13.03.23; peer-reviewed by A Wahbeh, N Yiannakoulis; comments to author 17.07.23; revised version received 30.07.23; accepted 16.12.23; published 29.01.24.

Please cite as:

Lu J, Zhang H, Xiao Y, Wang Y

An Environmental Uncertainty Perception Framework for Misinformation Detection and Spread Prediction in the COVID-19 Pandemic: Artificial Intelligence Approach

JMIR AI 2024;3:e47240

URL: <https://ai.jmir.org/2024/1/e47240>

doi: [10.2196/47240](https://doi.org/10.2196/47240)

PMID: [38875583](https://pubmed.ncbi.nlm.nih.gov/38875583/)

©Jiahui Lu, Huibin Zhang, Yi Xiao, Yingyu Wang. Originally published in JMIR AI (<https://ai.jmir.org>), 29.01.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR AI, is properly cited. The complete bibliographic information, a link to the original publication on <https://www.ai.jmir.org/>, as well as this copyright and license information must be included.

Original Paper

The “Magical Theory” of AI in Medicine: Thematic Narrative Analysis

Giorgia Lorenzini¹, MA, PhD; Laura Arbelaez Ossa¹, MD, PhD; Stephen Milford¹, DPhil; Bernice Simone Elger^{1,2}, DPhil; David Martin Shaw^{1,3*}, DPhil; Eva De Clercq^{1*}, DPhil

¹Institute for Biomedical Ethics, University of Basel, Basel, Switzerland

²Unit for Health Law and Humanitarian Medicine, Center for Legal Medicine, University of Geneva, Geneva, Switzerland

³Health, Ethics and Society, Universiteit Maastricht, Maastricht, Netherlands

*these authors contributed equally

Corresponding Author:

Giorgia Lorenzini, MA, PhD

Institute for Biomedical Ethics

University of Basel

Bernoullistrasse 28

Basel, 4056

Switzerland

Phone: 41 61 207 17 86

Email: giorgia.lorenzini@unibas.ch

Abstract

Background: The discourse surrounding medical artificial intelligence (AI) often focuses on narratives that either hype the technology’s potential or predict dystopian futures. AI narratives have a significant influence on the direction of research, funding, and public opinion and thus shape the future of medicine.

Objective: The paper aims to offer critical reflections on AI narratives, with a specific focus on medical AI, and to raise awareness as to how people working with medical AI talk about AI and discharge their “narrative responsibility.”

Methods: Qualitative semistructured interviews were conducted with 41 participants from different disciplines who were exposed to medical AI in their profession. The research represents a secondary analysis of data using a thematic narrative approach. The analysis resulted in 2 main themes, each with 2 other subthemes.

Results: Stories about the AI-physician interaction depicted either a competitive or collaborative relationship. Some participants argued that AI might replace physicians, as it performs better than physicians. However, others believed that physicians should not be replaced and that AI should rather assist and support physicians. The idea of excessive technological deferral and automation bias was discussed, highlighting the risk of “losing” decisional power. The possibility that AI could relieve physicians from burnout and allow them to spend more time with patients was also considered. Finally, a few participants reported an extremely optimistic account of medical AI, while the majority criticized this type of story. The latter lamented the existence of a “magical theory” of medical AI, identified with techno-solutionist positions.

Conclusions: Most of the participants reported a nuanced view of technology, recognizing both its benefits and challenges and avoiding polarized narratives. However, some participants did contribute to the hype surrounding medical AI, comparing it to human capabilities and depicting it as superior. Overall, the majority agreed that medical AI should assist rather than replace clinicians. The study concludes that a balanced narrative (that focuses on the technology’s present capabilities and limitations) is necessary to fully realize the potential of medical AI while avoiding unrealistic expectations and hype.

(JMIR AI 2024;3:e49795) doi:[10.2196/49795](https://doi.org/10.2196/49795)

KEYWORDS

artificial intelligence; medicine; physicians; hype; narratives; qualitative research

Introduction

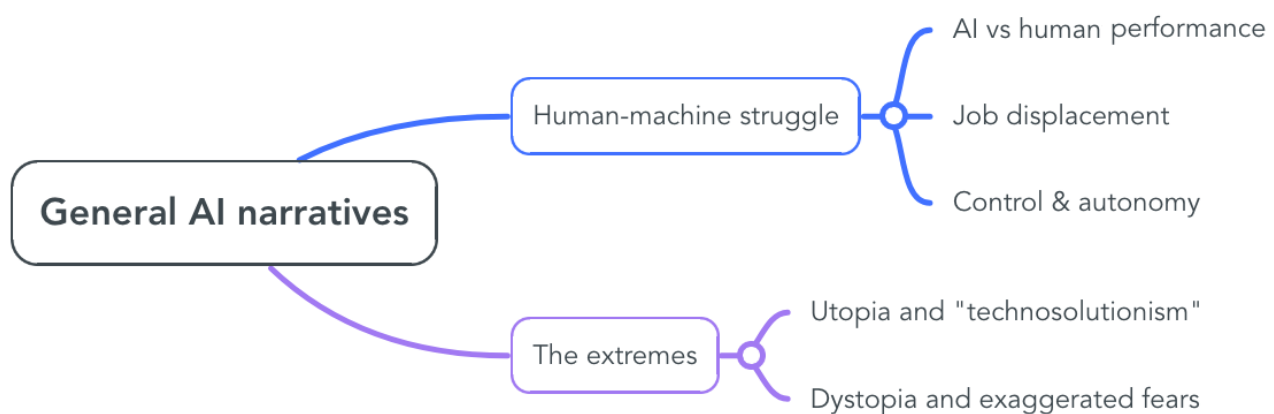
Background

Artificial intelligence (AI) technologies are steadily emerging and intertwining with humans' everyday lives and practices. Their applications are broad and diverse: in the field of health care, AI tools are supporting administrative tasks, predicting patients' prognoses, monitoring health through wearable devices, reading computed tomography scans, accelerating drug discovery and development, and many more applications [1]. Particularly relevant for the present analysis are AI-enabled wearable devices (eg, smartwatches) and clinical decision support systems (CDSSs). CDSSs are AI-based tools that provide diagnostic and treatment suggestions based on patient data and test results [2,3]. They bear the potential to impact physicians' clinical judgment, decision-making process, and their relationship with patients [4]. Lately, CDSSs are being combined with machine learning and deep learning techniques, thus generating hopes for faster and more accurate medical decisions and diagnoses [5]. Machine learning and deep learning are types of AI that continuously learn from the data they are fed [6]. Both wearables and CDSSs are artificial narrow intelligence as they are designed to perform only specific tasks. On the contrary, humans have general intelligence: they can excel in speech recognition, pattern recognition, decision-making, and creating. This is also the goal of AI research: with artificial general intelligence, the aim is to apply the same tool to different areas with similar satisfactory results and performance [7]. As artificial general intelligence is not currently a possibility, this paper focuses on artificial narrow intelligence applied in the medical context as CDSSs or wearable devices.

Our work rests on 2 pillars: the first is medical AI, and the second is the creation and perpetuation of AI narratives by people exposed to AI in their profession. It is in the nature of humans to make sense of things, events, and situations. One way of doing this is through the construction of narratives that link together complex and multifaceted realities while assigning roles, identities, and values. Narratives are, therefore, stories

we tell about our lives in a nuanced meaning-making effort [8]. It is important to analyze narratives because they reveal our attitudes, opinions, relationships, and emotions [9]. There is a multitude of general AI narratives (Figure 1), which come mainly from news outlets, science fiction accounts, the technology industry, and academic research. Prominent general AI narratives extensively concentrate on the struggle between humans and machines on different levels (ie, comparing their performances, worrying about job displacement, and wondering to which extent humans will relent control to AI). On the one hand, envisioning a world where AI takes over routine and tedious chores can be uplifting. On the other hand, it seems impossible to put to rest the underlying fear that it will take over everything else too, including more enjoyable and creative tasks [10]. Consequently, job displacement narratives are created based on the preoccupation that AI will render many jobs obsolete, particularly the ones revolving around menial tasks that could easily be automated [11]. This worry is exacerbated by the relentless comparison between humans' and AI's performances, as a means to validate AI's capabilities [12]. In this human-machine struggle, AI is depicted as a superefficient tool at the service of a heartless capitalistic system [10]. At the same time, AI is appreciated exactly because it holds the potential to simplify humans' lives: it is designed to help humans accomplish more with less effort. AI's achievements are often publicly praised; this is continuously underlined when its performance excels humans' capabilities. Accordingly, positive emotions and optimism are prevalent in social media posts about AI, also when the authors are experts in the field [13]. However, what is not acknowledged as much is that these successes are confined to very specific tasks: an AI that can excel in facial recognition will not automatically perform better than humans in driving cars. The lack of generalizability in AI means that human control and oversight are still pretty much needed. Having said that, narratives on AI taking control of human lives and societies are vastly popular [14]. What is usually incorrectly implied behind these narratives is that AI shares the human desire for greediness and its survival instincts, thus attributing these qualities to anthropomorphized machines [10,15].

Figure 1. Summary of general artificial intelligence (AI) narratives identified in the literature and pertinent to this analysis.



Dominant AI narratives are often mistrusted or criticized in light of their extremism: they frequently depict either utopian or dystopian futures, light-years away from the complex and

mundane reality, that misrepresent the present state of the technology [16]. For example, the way AI fails in the real world is far less epic and catastrophic from Hollywood conceptions:

these failures usually happen when the AI does what it is programmed to do but with unintended consequences, that is, a robot trained to behave in ways that would meet humans' approval pretending to be doing something useful [14,17]. The perpetuation of unrealistic AI failures inflates implausible fears while failing to address the real ways in which AI could fail [14]. The debate about AI is very polarized, and as opposed to apocalyptic predictions, there are overly optimistic accounts. The idea of AI being a "master technology" that would be able to unlock all sorts of useful technologies, including those that could help humanity achieve immortality, is common [18]. This leads to the imagination that AI could be considered a form of "holy grail" that bears the potential not only to provide for humanity's needs but also to fulfill its wildest desires and dreams [18].

Narratives can have different functions for different authors and in different situations; in this analysis, the focus is on how narratives could influence medical AI development and uptake and particularly how they could foster a climate where medical AI supports physicians. Indeed, narratives on AI have the power to influence the further development of these technologies, the availability of funding, the directions of research, and the opinions and expectations of both experts and the public. They influence how new sociotechnical realities are accepted and address both the concerns and the hopes surrounding AI [19]. Therefore, they form the background against which AI is being developed, interpreted, and assessed [16]. While general AI narratives are widely studied and debated, particularly in the Western world [14,19-21], little data are available on AI applications in specific sectors. The lack of research on medical AI narratives, coupled with the perception of AI being particularly promising in the field of health care [22,23], calls for more attention to the topic. Humans have a "narrative responsibility" [24]: there is a duty to make sense of medical AI and to do it responsibly because these sensemaking processes concretely impact its development, implementation, and uptake. Since the stories humans tell about medical AI shape the future of health care, narratives cannot be conceived as normatively neutral. Narratives that support how we wish medicine to be for the years to come should be preferred [8].

Objective

This paper offers a critical reflection on the existing literature on AI narratives. It is one of the first studies to examine the stories told by people who are professionally exposed to medical AI about its applications. This study compares these stories with the existing dominant general AI narratives so as to uncover meaningful similarities and differences. This study aims to raise awareness of how we talk about medical AI and how this can shape the future experiences of both patients and physicians. It is expected that some general AI narratives will be present in medical AI narratives. However, as this medical AI is implemented in a specific sector, namely, health care, with its particular features and challenges, some narratives will be unique for this context. The goal is to understand these similarities and differences to better evaluate medical AI narratives. Consequently, this study aims to recommend a more ethical approach when creating and perpetuating these

narratives, considering their impact on physicians' jobs and the physician-patient relationship.

Methods

Overview

The data used for this manuscript are part of a larger research project titled Ethical and Legal issues of Mobile Health-Data: Improving Understanding and Explainability of Digital Transformation and Data Technologies Using Artificial Intelligence (EXPLaiN), which aims to clarify the legal and ethical issues that need to be resolved for the collection, use, and analysis of health data with AI methods. The project is funded by the Swiss National Science Foundation. The first part of the study consisted of 41 semistructured interviews with participants who are exposed to medical AI. These participants were from a range of disciplines: medicine, philosophy, law, ethics, public health, and computer science. The interviews focused on the barriers and facilitators for the implementation of AI in clinical settings, particularly regarding CDSSs and wearable devices. The original study aimed to examine the current views, attitudes, knowledge, and barriers to using AI models in the analysis of health data and to support physicians and patients in their decision-making.

This analysis is a secondary analysis of these data and focuses on a subset of the data collected. While coding the data, it became apparent that narratives were often discussed. This justified a secondary analysis that was attentive to this aspect of the data. A second code tree was created based on the narratives identified in the literature, and the interviews were recoded. Of the 41 interviews, 30 (73%) were selected for the secondary analysis based on the presence of narrative elements about AI in health care. This selection, inherent to the secondary nature of the analysis, resulted in incomplete saturation in 1 subtheme, namely, "welcoming the holy grail."

The data subset was analyzed using a thematic narrative approach that identified and reported stories participants told about medical AI [25,26]. This approach was chosen for its flexibility and ability to allow large data sets to be managed and reduced into themes [25]. The topic and the format of the data are not conducive to a structural narrative approach, as the narrative segments were relatively short and lacked common narrative characteristics (eg, characters with roles, a narrator, a complication, a resolution, and a coda) [27,28]. Therefore, a narrative thematic analysis was chosen, as it enabled single units of meaning, primarily phrases, and short paragraphs to be formed into themes and interpreted narratively [29]. With a narrative thematic approach, we could better describe how people exposed to AI in their profession experienced and understood medical AI, as well as how they made sense of it. This allowed for the analysis of underlying assumptions and values [27,30].

Participants

Participants were purposively sampled and came from various disciplines and backgrounds: medicine, bioethics, public health, philosophy, psychology, economy, law, and computer science. Inclusion criteria, other than being exposed to medical AI in

their profession, were the holding of a senior position, either in academia or in the private sector, hence excluding PhD students, interns, and junior professionals. Participants' profiles were, for example, full professorship at a university, chief executive officer of a company working with AI, or a data protection officer at a hospital. Participants were recruited internationally; however, there was a focus on European and Swiss participants since the EXPLaiN project aimed to especially explore their attitudes. Participants were recruited because they were working with medical AI through projects, products, research, and development. Identification of participants occurred through publications or affiliations with companies working in the field of medical AI. Their email addresses were found on the web through their institution or company's website. At the end of the interview, participants were asked if they knew someone meeting the inclusion criteria who would be interested in participating (snowball sampling).

First contact with participants was through email where they were invited to be interviewed by introducing the project and explaining the aims and the implications of their participation (eg, time commitment, voice recording, the method of transcription, and data pseudonymization format).

Data Collection and Analysis

LAO and GL, who recruited the participants and conducted the one-on-one semistructured interviews, did not personally know the participants. LAO has a background in medicine and public health, while GL studied philosophy with a focus on ethics and

philosophy of science. At the time of the data collection, both were PhD candidates in bioethics at the Institute of Biomedical Ethics of Basel.

Data were collected from November 2021 to April 2022 (therefore preceding some breakthrough such as ChatGPT; it could be hypothesized that after the most recent novelties in the AI field, such as natural language processing tools, narratives about AI might be different also in the health care sector) through semistructured interviews that lasted an average of 50 minutes. All the interviews of this subset were conducted on the web and recorded directly via Zoom (Zoom Video Communication, Inc). The original interview guide was composed of 13 questions, each with several prompts or follow-ups. The interview guide made use of 3 vignettes to better clarify and contextualize the questions. The questions were divided into 6 blocks: introductory questions (about the experience of the participant), general questions about using AI in medical practice, context-related questions about AI-patient relationships (vignette 1 involving a wearable device), context-related questions about physician-patient relationships with AI (vignette 2 involving CDSSs), context-related questions about private-public relationships (vignette 3), and closing questions. The more significant questions (reported in [Textbox 1](#)) for this analysis were questions numbered 3 and 4, as well as 3 prompts for question 8. However, relevant data were found elsewhere in the data set as the interviews were semistructured, and participants had some freedom in guiding the topics of the interview.

Textbox 1. Relevant questions and prompts from the interview guide.

Question numbers and questions

- 3: I would like to start discussing clinical usability. What do you think about using artificial intelligence (AI) in clinical practice?
- 4: What would you consider the biggest challenges of using AI in health care?
- 8.6: How important it is that the physician understands AI?
- 8.8: Would AI have an impact on the physician-patient relationship?
- 8.9: Would AI challenge the traditional model of shared decision-making?

The interviews were transcribed verbatim by LAO, GL, and 2 students at the University of Basel using MAXQDA (VERBI GmbH), a software application designed to assist with qualitative analysis methods. LAO and GL checked all the transcripts and compared their correctness with the audio of the interviews. All data were securely stored on the server of the University of Basel and pseudonymized. Potentially reidentifiable information was removed from the transcripts.

After the original inductive coding, conducted equally by GL and LAO, GL reorganized the relevant coded sections for the secondary analysis. Upon consulting existing literature to identify dominant narratives, a new code tree was created, and the selected segments were deductively recoded. The selected data subset was interpreted through the lens of the existing categories of general AI narratives [31]. The new code tree composed the dominant AI narratives found in the literature. GL then selected the most significant codes and grouped them into themes.

Ethical Considerations

All methods were approved by the Ethics Committee of Northwest and Central Switzerland, under Switzerland's Human Research Act (HRA) Article 51 [32]. The methods were carried out in accordance with the relevant HRA guidelines and regulations. After revision, the Ethics Committee of Northwest and Central Switzerland concluded that interviewing AI professionals falls outside the HRA and requires only verbal consent at the beginning of an interview (declaration of no objection AO_2021-00045).

All personal data were pseudonymized and safely stored on the server at University of Basel. The key is accessible only to the research team. Potentially reidentifiable data were omitted from publication. No compensation was offered to participants.

Results

Overview

For this analysis, we used 30 interviews and reported at least 1 quote from each. We selected this subset because narratives were not prominent in all interviews. It was challenging to categorize participants into disciplines, as AI is notoriously an interdisciplinary field. More often than not, participants had mixed backgrounds and were dealing with medical AI from different points of view. In categorizing participants, we picked their main expertise: 9 (30%) participants had a background in medicine, 6 (20%) in bioethics, 6 (20%) in law, 3 (10%) in computer science, 2 (7%) in public health, 2 (7%) in philosophy,

1 (3%) in psychology, and 1 (3%) in economy. The vast majority of the selected participants (21/30, 70%) were male (female participants: n=9, 30%). Only 5 (17%) participants were located outside Europe: 3 (10%) in the United States, 1 (3%) in Canada, and 1 (3%) in South Africa (for more details on the participants, please refer to the [Multimedia Appendix 1](#)).

Our analysis identified 2 main themes and various subthemes ([Figure 2](#)). Representative anonymized quotes were taken from the interviews to illustrate the reported results. Participants are identified with the abbreviation of their main expertise and a number: bioethics (BE), computer science (CS), economy (EC), law (LW), medicine (ME), public health (PH), philosophy (PL), and psychology (PS).

Figure 2. Themes and subthemes that emerged from the thematic narrative analysis. AI: artificial intelligence; MAI: medical artificial intelligence.



Medical AI as a Game Changer

Overview

With regard to physicians and medical AI relationships, attitudes fell into 2 main groups. Some participants depicted a rather competitive relationship and compared the performances and capabilities of medical AI to those of physicians. The majority, however, emphasized how AI can support clinicians, thus outlining a more collaborative relationship and focusing on the benefits of this cooperation. Nevertheless, these 2 groups shared the underlying idea that AI would be a game changer for medicine, and both emphasized how it could be useful in health care.

A Competitive Relationship

Medical AI Competing With Physicians

Some participants described AI as a competitor to physicians and argued that not only clinicians are dependent on AI but also they could even be replaced by it. Medical AI was said to outperform clinicians in pattern recognition and data processing. AI was believed to notice aspects that physicians would miss, hence emphasizing the limitations of human capabilities and describing AI as being faster, more accurate, and less costly than human physicians:

AI is able to grasp so many ideas within a very small time interval [and] also integrate information that physicians might even oversee that this might be even like more precise than physicians. And I think this is also an advantage. [ME7]

The AI tool uncovers a pattern that the clinician did not pick up or maybe could not have picked up within a human limited abilities. [BE5]

[Medical AI is] very inexpensive to use. In principle, like once you've trained the system for let's say a diagnosis, you can basically use these things on a regular laptop or smartphone even..... It doesn't come for free, but it is rather inexpensive and easy to get. [BE4]

Comparing physicians and AI performance, abilities, and costs sometimes resulted in claims about the obsolescence of physicians since AI would be better in many aspects of a physician's role, while also being faster and cheaper, and it seemed to be preferable to delegating tasks to AI. It was rarely implied that physicians as a whole would be replaced. More commonly, it was suggested that some specific tasks could be carried out by AI. A common limitation was that AI could not interact with patients as at present it lacked the necessary skills. Presuming that AI capacities would steadily improve, a few participants wondered whether in the future medical AI might be able to assume all physicians' duties:

Nowadays there are certain things that might not be outsourced to machines in terms of human interactions. But on the other side, I think, if we wait long enough, we can basically outsource everything to machines. [PH1]

I'm pretty sure that the physician will be quite cautious, at least in the beginning, when they know that they use these kinds of products [medical AI], but maybe with time, you know, when they are used

to it, in like 5 years, 10 years, 15 years, maybe with time they could lose probably autonomy. [LW6]

You will actually have better outcomes if you don't involve humans. [ME9]

Things were different for image recognition: several participants mentioned that medical AI gave outstanding results in radiology. This led to the possibility of outsourcing routine cases to AI while consulting radiologists only for peculiar cases:

I think one of the big places where it's already implemented is in radiology. Meaning, recognition of patterns in pictures; machines are better at it than we are. [ME1]

What people have been doing in radiology, I think it's also awesome.... The machine can give you feedback right away and maybe you just use the humans for very specific cases. [CS1]

And as more people use that tool, there might be the temptation that therefore maybe we don't need as many dermatologists. Or as many specialists in certain areas, like radiology. Because we have very good AI that is able to detect cancer from X-rays. Or covid from X-rays of lungs. [PS1]

The Risk of Technological Deferral

While pondering the idea of a more autonomous medical AI, many participants worried about the risks of excessive technological deferral (giving too much power to technology). Automation bias, namely, the tendency to overrely on automatic decision-making tools, was mentioned as an issue in areas of practice that are time-sensitive:

In the long run [physicians] end up with them just following what the machine says. [PL1]

There is a very real risk, especially in areas of practice that have time pressure, that we will see automation bias, that we will see AI systems that formally were advisory, actually being the ones who decide treatment choices. [BE2]

This tension on who holds the final decision-making responsibility was framed as an actual conflict, with potentially detrimental consequences if the humans were to “lose” their decision-making power. Physicians might also be intimidated by this outstanding tool and therefore would struggle to override its decisions even when they did not agree with it:

Can you even win, so to speak? So, that might be the bigger danger, where you say like “well, the machine says that, so therefore it is correct.” [PL2]

The recurrent mentioned consequences of deferring decision-making powers to medical AI were dependency on technology, with fewer and fewer specialists trained and a gradual loss of autonomy for physicians. Many participants in this group worried about physicians' autonomy being endangered by medical AI and described the technology as authoritarian or tyrannical:

Well, if the algorithms prove to be better than physicians then you would have to change the role of

physicians from decision-makers to more just like people, in the end, giving injections. [CS2]

A Collaborative Relationship

The Question of Irreplaceability

For many participants, physicians are not to be replaced by AI; rather, AI enables them and supports their daily practice and decision-making activities:

It should go in the direction that the systems are not seen as a competitor to the physicians but more as a cooperation between both. And I think what it's worthwhile, what it's important, it's that the cooperation leads to better results. [EC1]

The use of technology is going to assist the physician and not harm because in the end it's called a clinical decision support tool, not a clinical decision maker tool. [ME3]

But I think it will never, never replace the main diagnosis of a physician. So, this will always be a support tool. Which has to be as well validated beforehand. [LW2]

We need to be clear that AI is not just coming along to replace physicians and when they go to the GP [general practitioner], they're going to see a robot instead and the robot won't understand anything about them and it'll just give them a stamp prescription that is the same as everyone else. That's not what AI is. And certainly not in the next few decades, will it be used for anything other than decision support. [ME5]

Some interviewees noted how humanity is irreplaceable, while others described medical AI as an assistive tool that is not designed to replace physicians but to empower them. Participants in this group emphasized the idea of medical AI “assisting,” “helping,” “empowering,” and “supporting” clinicians rather than comparing their ability, accuracy, and cost.

When emphasizing physicians' irreplaceability, participants referred to the sensibility, emotivity, and empathy that are needed in medical decision-making. Given the current state of the art, medical AI is unable to grasp the complex totality of the patient's situation. Many participants also questioned patients' willingness to relinquish the physician-patient relationship in favor of an AI-patient relationship. They argued that communication with AI would not be authentic, as it would not consider patients' personhood. Therefore, these participants concluded that medical AI should never override physicians' decisions; rather, it should promote and preserve physicians' agency:

The patient needs a person he can talk [to], a person that can read their emotions, feelings. [LW4]

I think medicine has a certain degree of nuances, that only a person might catch and not a computer. And you can't let these computers or AI run autonomously. [ME4]

Independently from AI capabilities (whether it outperforms physicians or not and whether it is limited or not), physicians remain essential: medical AI should always be considered in the light of physicians' clinical judgment and never left unsupervised. According to this description of medical AI, physicians should always keep an active role in decision-making:

I think the one who has the responsibility to make decisions is, or will always be, the physician. [LW4]

I expect that the technology will help you give an assessment, but that you will still have a clinician that will evaluate further that kind of technical assessment by software. So it's not fully replacing an intervention as such. It's helping, supporting a development. [LW5]

I think that the physician has to make a decision based on their training. That's their responsibility. [ME8]

Medical AI Freeing Physicians

The collaborative relationship narrative does not depict physicians as dependent on their tools; rather, it suggests that medical AI could constitute an important resource. The relationship is described as a fruitful partnership, and the outcome would be a general improvement to both physicians' practice and their work conditions. Medical AI could free physicians from burdensome tasks, hence relieving them from burnout and allowing them to spend more quality time with patients:

[Medical AI] could improve the physician-patient communication.... So I am kind of hoping that, in that way, because of AI certain aspects of healthcare could be simplified and automated, but that equally should generate room for more empathy between physicians and patients. [PS1]

[Medical AI] is helping physicians to really focus on, or be able to have more time for patients and less to spend with tools. [CS3]

What I hope it'll do it's improve the relationship between the patient and the physician. What I mean by this is [that] the physician is going to be relieved from the burnout. [ME3]

The Power of Medical AI

Overview

Most of the participants were optimistic about the future of medicine when AI was involved and reported an overall positive impact, or potential, of this technology. While a large part of the answers balanced medical AI's advantages with the challenges it introduces, some focused only on the benefits of the technology. At the same time, many interviewees identified a hype-type narrative of medical AI and problematized it. In this context, hype is understood as an exaggeratedly optimistic rhetoric about an emerging technology [33].

Welcoming the Holy Grail

In a few interviews, medical AI was discussed mainly in positive terms. These participants did not see any negative aspects or

concerns about the technology. Medical AI was deemed always useful, and if it was not useful for something yet, it surely would be in the future. It encapsulated so many opportunities for health care that 1 participant referred to it as "the holy grail." Consequently, medical AI was expected to solve a wide range of problems:

I basically don't see any negative effects, like, I can't really see any negative effects. [LW1]

So, it seems to me that it's both inevitable and good that we have it [medical AI]. [BE2]

What do you think about using AI or machine learning in clinical practice? [GL] I think it's the Holy Grail. [CS1]

Well, it [medical AI]'s a game changer. And I think that our wild dream about getting personalized medicine is really at hand. [LW3]

Medical AI Is Not Magical

A significant part of the participants addressed the romanticization of this technology and highlighted the importance of promoting a more truthful narrative. "Truth" and "reality" were terms often mentioned when discussing the medical AI hype: it was deemed untruthful, unhelpful, and unrealistic, and this was judged problematic:

The problem is that this enthusiasm is so uncritical and then we build into this. This is not giving us the truth and not helping us to generate probabilities. This is the problem that I hugely see. [BE3]

According to the participants, one of the consequences of the hype around medical AI is that it is impossible to live up to the expectations that it builds. Therefore, some participants were profoundly critical toward overhyped accounts of the capabilities of medical AI:

There is so much hype in this field [medical AI] and this builds narratives and expectations. And to live up to those expectations is always challenging. [ME2]

So that has, probably now backwards looking, not been so clever to phrase it as the silver bullet solution to everything, to patient autonomy, or patient empowerment, to more efficient and better healthcare. [BE1]

The outcome of this ideology is that medical AI is portrayed as the appropriate means to tackle every pressing issue of health care: AI is the hammer that fixes everything. Techno-solutionist narratives would misunderstand AI and promote a representation of the technology as if it were some kind of magical tool:

The hype around the technology at the moment, you know, that people think that it can solve everything. It's like they have a hammer and everything is a nail. [PH2]

I think a lot of people and a lot of physicians kind of have the magical theory of machine learning, where you just kind of throw the numbers in the hopper, shake it out, and you get the results by magic. [BE6]

The major problem of deep learning today are the people doing deep learning because they think they will solve everything with that and the ignorant people because they don't understand what is deep learning and they think it's magic that will solve everything.
[ME6]

Discussion

Principal Findings

The accounts of medical AI that emerged from these interviews are more realistic and less influenced by science fiction narratives than the general discourse on AI. Dystopian futures were not reported, and only a few participants described AI as a utopian technology that would address all challenges faced by the health care sector. While general AI narratives are usually polarized, describing AI as either the milestone of a better future for humanity or the cause of all evils [19,34,35], study participants often found a middle path between medical AI's promises and risks, thus avoiding alignment with extreme positions and providing instead a more nuanced depiction of the technology. We hypothesize that people exposed to AI in their profession are less prone to exaggerated and polarized narratives, while lay people tend to be more susceptible to these narratives as they feel they have less control over the technology [36]. The lack of a strongly polarized discourse in medical AI can be regarded as positive: the contradictions present in narratives that are diametrically opposed and irreconcilable hinder a nuanced and sophisticated understanding of the technology [21].

However, our study sample was not exempt from hype narratives that uncritically focused on the expected benefits of medical AI. This confirmed the existence of hype narratives, which are already reported in the literature as well as the conceptualization of AI as a "holy grail" technology [22,23,34].

Claims about superiority are very popular in AI narratives, not only in fictional and media narratives but also in the scientific discourse, as researchers frequently compare AI with humans' capabilities and performances as a means of validating the technology [12,37]. The physician-AI juxtaposition ends with depicting the classical human-machine struggle panorama, where physicians are menaced by an authoritarian machine that outperforms them and that leaves humans dependent on it, no longer in control, and stripped of their agency [12,14,19,35]. Indeed, 1 participant described this struggle as a real win-lose situation.

While a few participants hyped medical AI, the majority recognized both the advantages and the challenges introduced by AI in health care. Therefore, stronger than the hype narrative were the cautionary tales of avoiding a "myopic techno-solutionism" and the criticism of this hype [34]. Techno-solutionism is the ideology in which every kind of problem (technical, social, economic, political, psychological, or physical) can be ameliorated with an "appropriately designed" technological solution [38]. Attributing magical properties to AI, meaning that it can somehow address every problem, reveals a shallow understanding of the technology. This requires better education, which can be achieved through the establishment of

a more balanced narrative that realistically assesses medical AI's current capabilities and shortcomings [37].

Participants confirmed the idea that medical AI narratives can sometimes be detached from the everyday reality of the technology and that the hyping of AI leads to unrealistic expectations and overpromising while obscuring technological bottlenecks [19,21]. Therefore, our findings demonstrate that the current dominant narratives can mislead the understandings of medical AI, even in people working with it. Instead, "narratives should focus on the realities of AI's present capabilities" [34] and take into account the narrative responsibility that is always entailed when the future of medicine with AI is imagined. Every story we tell about medical AI shapes its development, adoption, and perception in health care in ways that are not normatively neutral.

Accordingly, almost all participants recognized the limitations of AI. There is a risk that by failing to acknowledge the potential problems and shortcomings of medical AI, the hype narrative might further exacerbate these hidden specters. The need for a more realistic narrative that returns the image of the actual state of the art is commonly present both in the interviews and in the debate about AI narratives [14,19,34].

With the exception of a few participants, there was a general agreement that AI could not and would not replace human clinicians. This finding is present in the literature about the future of medicine with AI; for example, patients appeared less prone to seek medical assistance if AI provides it, even if it was better than a human expert [39]. When it comes to this topic, there is an alignment between different narratives that appear to share similar moral codes according to which medical AI cannot entirely replace the physician's role or human interaction [40]. Therefore, this could be regarded as the "proper narrative" of the AI-physician relationship, and, as such, it might take the form of a collective narrative or "imaginary," judged true without a need for further justification [41]. The prevailing idea remains: "patients will always need human physicians" [42].

Having determined that medical AI is to assist clinicians, it remains to be assessed whether it will have an impact on the physician-patient relationship. Some participants believed that medical AI would ameliorate their relationships, for example, by allowing physicians to spend more time with patients. This is also a popular idea in the literature to the extent that some claim that medical AI could be an opportunity to make physicians more human and empathetic [43-47]. However, as with many things about AI, opinions are divided, and this idea is also widely criticized. It could be that physicians will visit more patients in the time AI saved, thus maintaining the status quo or worsening care provisions [12,48,49]. Consequently, medical AI might not necessarily have a positive impact on the physician-patient relationship as either the participants in our study or many prominent voices in research think.

Limitations

There is a clear prevalence of a Western perspective in our study. Hence, it remains questionable whether our findings are valid in other contexts.

The interview guide that we used for this study focused on certain applications of AI in medicine, namely, CDSSs and wearable devices (eg, smartwatches). This may have limited the discourse on possible outcomes and futures. Moreover, question 8.8. discusses the “traditional” model of shared decision-making; this wording could be considered nonneutral and leading.

Before commencing this analysis, we have conducted theoretical research on the ethical issues of medical AI. This led us to publications where we took a position on the role of AI in health care and the physician-patient relationship. We concluded that medical AI is currently, and should continue to be, an assistive tool that should support physicians’ and patients’ decision-making. We acknowledge that this belief was already sedimented at the time of data collection and analysis, thus possibly shaping the way in which we presented the results.

Conclusions

Through the establishment of a more realistic and nuanced medical AI narrative, it is easier to describe AI tools as assistive. The discourse about their benefits, risks, and possible applications is less spectacular. Narratives that support the idea of AI augmenting humans’ capabilities, rather than substituting them, should be preferred as these narratives better correspond to the current reality of the technology [34]. It is also fundamental to raise awareness of the narrative responsibility that humans have to make sense of, interpret, and narrate medical AI in a way that shapes a positive future for medicine.

Similarly, humans are responsible for scrutinizing the dominant narratives and evaluating them [24]. Everyone has this responsibility when talking about medical AI, including researchers, since we all can impact the future of technology, although to different degrees. Failing to exercise this narrative responsibility would entail relinquishing our sense-making task to other narrators (eg, big tech, transhumanists, governments, etc). The consequence would be a world in which we live in the narrative created by others for us. This world would be one in which the majority of humanity delegated the construction of our future to a few, in that they did not participate in the process that would shape what mattered most in the present [24,50].

Disproportionate fears and expectations could halt the development of medical AI, for example, by generating opposition or disillusionment when the technology does not live up to its promised expectations [19,21]. Medical AI narratives shape the role of AI in societies in ways that are ethically and politically relevant and can influence the perceptions of citizens, policy makers, politicians, health care personnel, and researchers [8,16]. Therefore, narratives have a constitutive role that is more than strictly descriptive: it is performative. Narratives have the power to decide the future of medical AI [51,52]. We argue that it is important to recognize the role that narratives of technologies play for humanity and reflect on which type of narrative is dominant in medical AI. This is a fundamental ethical issue that cannot be overlooked. It must be addressed so as to shape our desired future for medicine.

Acknowledgments

The authors thank the people at the Institute for Biomedical Ethics of Basel for their support of this study, particularly, Dr Tenzin Wangmo and Dr Michael Rost for their precious expertise and kind advice. This paper and the research on which it is based are funded by the Swiss National Science Foundation (SNF) in the context of the Ethical and Legal issues of Mobile Health-Data: Improving Understanding and Explainability of Digital Transformation and Data Technologies Using Artificial Intelligence (EXPLaiN) project (National Research Projects 77; grant or award number 407740_187263/1).

Data Availability

The data sets generated during and analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions

GL prepared the original draft for this paper. GL, EDQ, and DMS participated in the conceptualization, in the methodology choice, and in the secondary data analysis. GL and LAO conducted the data collection and together with SM, and DMS participated in the original data analysis. BSE was responsible for funding acquisition. All the authors contributed to the review and editing of the article.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Participants’ characteristics.

[DOCX File , 15 KB - [ai_v3i1e49795_app1.docx](#)]

References

1. Basu K, Sinha R, Ong A, Basu T. Artificial intelligence: how is it changing medical sciences and its future? *Indian J Dermatol* 2020;65(5):365-370 [FREE Full text] [doi: [10.4103/ijid.IJD_421_20](#)] [Medline: [33165420](#)]

2. Sutton RT, Pincock D, Baumgart DC, Sadowski DC, Fedorak RN, Kroeker KI. An overview of clinical decision support systems: benefits, risks, and strategies for success. *NPJ Digit Med* 2020 Feb 6;3:17 [FREE Full text] [doi: [10.1038/s41746-020-0221-y](https://doi.org/10.1038/s41746-020-0221-y)] [Medline: [32047862](https://pubmed.ncbi.nlm.nih.gov/32047862/)]
3. Berner ES, La Lande TJ. Overview of clinical decision support systems. In: Berner E, editor. *Clinical Decision Support Systems: Theory and Practice*. Cham, Switzerland: Springer; 2007:1-17.
4. Lorenzini G, Arbelaez Ossa L, Shaw DM, Elger BS. Artificial intelligence and the doctor-patient relationship expanding the paradigm of shared decision making. *Bioethics* 2023 Jun 25;37(5):424-429. [doi: [10.1111/bioe.13158](https://doi.org/10.1111/bioe.13158)] [Medline: [36964989](https://pubmed.ncbi.nlm.nih.gov/36964989/)]
5. Wang D, Wang L, Zhang Z, Wang D, Zhu H, Gao Y, et al. “Brilliant AI doctor” in rural clinics: challenges in AI-powered clinical decision support system deployment. In: *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 2021 Presented at: CHI '21; May 8-13, 2021; Yokohama, Japan. [doi: [10.1145/3411764.3445432](https://doi.org/10.1145/3411764.3445432)]
6. Du Y, McNestry C, Wei L, Antoniadi AM, McAuliffe FM, Mooney C. Machine learning-based clinical decision support systems for pregnancy care: a systematic review. *Int J Med Inform* 2023 May;173:105040 [FREE Full text] [doi: [10.1016/j.ijmedinf.2023.105040](https://doi.org/10.1016/j.ijmedinf.2023.105040)] [Medline: [36907027](https://pubmed.ncbi.nlm.nih.gov/36907027/)]
7. Kaplan A, Haenlein M. Siri, Siri, in my hand: who’s the fairest in the land? On the interpretations, illustrations, and implications of artificial intelligence. *Bus Horiz* 2019 Jan;62(1):15-25 [FREE Full text] [doi: [10.1016/j.bushor.2018.08.004](https://doi.org/10.1016/j.bushor.2018.08.004)]
8. Coeckelbergh M. Time machines: artificial intelligence, process, and narrative. *Philos Technol* 2021 Oct 23;34:1623-1638. [doi: [10.1007/s13347-021-00479-y](https://doi.org/10.1007/s13347-021-00479-y)]
9. Wertz FJ, Charmaz K, McMullen LM, Josselson R, Anderson R, McSpadden E. *Five Ways of Doing Qualitative Analysis: Phenomenological Psychology, Grounded Theory, Discourse Analysis, Narrative Research, and Intuitive*. New York, NY: Guilford Publications; 2011.
10. Dihal K. Enslaved minds: artificial intelligence, slavery, and revolt get access arrow. In: *AI Narratives: A History of Imaginative Thinking about Intelligent Machines*. Oxford, UK: Oxford University Press; 2020:189-212.
11. Guenduez AA, Mettler T. Strategically constructed narratives on artificial intelligence: what stories are told in governmental artificial intelligence policies? *Gov Inf Q* 2023 Jan;40(1):101719. [doi: [10.1016/j.giq.2022.101719](https://doi.org/10.1016/j.giq.2022.101719)]
12. Ostherr K. Artificial intelligence and medical humanities. *J Med Humanit* 2022 Jun 11;43(2):211-232 [FREE Full text] [doi: [10.1007/s10912-020-09636-4](https://doi.org/10.1007/s10912-020-09636-4)] [Medline: [32654043](https://pubmed.ncbi.nlm.nih.gov/32654043/)]
13. Manikonda L, Kambhampati S. Tweeting AI: perceptions of lay versus expert Twitterati. In: *Proceedings of the International AAI Conference on Web and Social Media*. 2017 Presented at: ICWSM-17; May 15-18, 2017; Montreal, Quebec. [doi: [10.1609/icwsm.v12i1.15061](https://doi.org/10.1609/icwsm.v12i1.15061)]
14. Recchia G. The fall and rise of AI: investigating AI narratives with computational methods. In: Cave S, Dihal K, Dillon S, editors. *AI Narratives: A History of Imaginative Thinking about Intelligent Machines*. Oxford, UK: Oxford University Press; 2020:382-408.
15. Singler B. Artificial intelligence and the parent–child narrative. In: Cave S, Dihal K, Dillon S, editors. *AI Narratives: A History of Imaginative Thinking about Intelligent Machines*. Oxford, UK: Oxford University Press; 2020.
16. Cave S, Dihal K, Dillon S. Introduction: imagining AI. In: Cave S, Dihal K, Dillon S, editors. *AI Narratives: A History of Imaginative Thinking about Intelligent Machines*. Oxford, UK: Oxford University Press; 2020.
17. Christiano P, Leike J, Brown TB, Martic M, Legg S, Amodei A. Deep reinforcement learning from human preferences. Preprint posted online June 12, 2017 [FREE Full text]
18. Cave S. AI: artificial immortality and narratives of mind uploading. In: Cave S, Dihal K, Dillon S, editors. *AI Narratives: A History of Imaginative Thinking about Intelligent Machines*. Oxford, UK: Oxford University Press; 2020.
19. Cave S, Craig C, Dihal K, Dillon S, Montgomery J, Singler B, et al. Portrayals and perceptions of AI and why they matter. *The Royal Society*. 2018 Nov. URL: <https://royalsociety.org/-/media/policy/projects/ai-narratives/ai-narratives-workshop-findings.pdf> [accessed 2024-07-23]
20. Cave S, Dihal K. The whiteness of AI. *Philos Technol* 2020 Aug 06;33:685-703. [doi: [10.1007/s13347-020-00415-6](https://doi.org/10.1007/s13347-020-00415-6)]
21. Vicsek L. Artificial intelligence and the future of work – lessons from the sociology of expectations. *Int J Sociol Soc Policy* 2020 Oct 06;41(7/8):842-861. [doi: [10.1108/ijssp-05-2020-0174](https://doi.org/10.1108/ijssp-05-2020-0174)]
22. Cameron D, Maguire K. Public views of machine learning: digital natives. *The Royal Society*. 2017 Oct. URL: <https://royalsociety.org/-/media/policy/projects/machine-learning/digital-natives-16-10-2017.pdf> [accessed 2024-07-23]
23. Frost EK, Carter SM. Reporting of screening and diagnostic AI rarely acknowledges ethical, legal, and social implications: a mass media frame analysis. *BMC Med Inform Decis Mak* 2020 Dec 10;20(1):325 [FREE Full text] [doi: [10.1186/s12911-020-01353-1](https://doi.org/10.1186/s12911-020-01353-1)] [Medline: [33302942](https://pubmed.ncbi.nlm.nih.gov/33302942/)]
24. Coeckelbergh M. Narrative responsibility and artificial intelligence: how AI challenges human responsibility and sense-making. *AI Soc* 2021 Dec 30;38(6):2437-2450. [doi: [10.1007/s00146-021-01375-x](https://doi.org/10.1007/s00146-021-01375-x)]
25. McAllum K, Fox S, Simpson M, Unson C. A comparative tale of two methods: how thematic and narrative analyses author the data story differently. *Commun Res Pract* 2019 Nov 25;5(4):358-375. [doi: [10.1080/22041451.2019.1677068](https://doi.org/10.1080/22041451.2019.1677068)]
26. Braun V, Clarke V. Using thematic analysis in psychology. *Qual Res Psychol* 2006 Jan;3(2):77-101. [doi: [10.1191/1478088706qp063oa](https://doi.org/10.1191/1478088706qp063oa)]

27. Bruner J. The narrative construction of reality. In: Beilin H, Pufall PB, editors. *Piaget's Theory: Prospects and Possibilities*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc; 1992:229-248.
28. Clarke V, Braun V. Thematic analysis. *J Posit Psychol* 2016 Dec 09;12(3):297-298. [doi: [10.1080/17439760.2016.1262613](https://doi.org/10.1080/17439760.2016.1262613)]
29. Ross JA, Green C. Inside the experience of anorexia nervosa: a narrative thematic analysis. *Couns Psychother Res* 2010 Jul 22;11(2):112-119. [doi: [10.1080/14733145.2010.486864](https://doi.org/10.1080/14733145.2010.486864)]
30. Fisher WR. Narration as a human communication paradigm: the case of public moral argument. *Commun Monogr* 2009 Jun 02;51(1):1-22. [doi: [10.1080/03637758409390180](https://doi.org/10.1080/03637758409390180)]
31. Braun V, Clarke V. Conceptual and design thinking for thematic analysis. *Qual Psychol* 2022 Feb 13;9(1):3-26. [doi: [10.1037/qup0000196](https://doi.org/10.1037/qup0000196)]
32. Federal Act on research involving human beings (Human Research Act, HRA), chapter 9: research ethics committees. Fedlex. 2011. URL: https://www.fedlex.admin.ch/eli/cc/2013/617/en#chap_9 [accessed 2024-07-23]
33. van Lente H, Spitters C, Peine A. Comparing technological hype cycles: towards a theory. *Technol Forecast Soc Change* 2013 Oct;80(8):1615-1628. [doi: [10.1016/j.techfore.2012.12.004](https://doi.org/10.1016/j.techfore.2012.12.004)]
34. Chubb J, Reed D, Cowling P. Expert views about missing AI narratives: is there an AI story crisis? *AI Soc* 2022 Aug 25:1-20 (forthcoming) [FREE Full text] [doi: [10.1007/s00146-022-01548-2](https://doi.org/10.1007/s00146-022-01548-2)] [Medline: [36039046](https://pubmed.ncbi.nlm.nih.gov/36039046/)]
35. Klarmann N. Artificial intelligence narratives: an objective perspective on current developments. Preprint posted online March 18, 2021 [FREE Full text]
36. Borup M, Brown N, Konrad K, Van Lente H. The sociology of expectations in science and technology. *Technol Anal Strateg Manag* 2006 Jul;18(3-4):285-298. [doi: [10.1080/09537320600777002](https://doi.org/10.1080/09537320600777002)]
37. Park SH, Do KH, Kim S, Park JH, Lim YS. What should medical students know about artificial intelligence in medicine? *J Educ Eval Health Prof* 2019 Jul 03;16:18 [FREE Full text] [doi: [10.3352/jeehp.2019.16.18](https://doi.org/10.3352/jeehp.2019.16.18)] [Medline: [31319450](https://pubmed.ncbi.nlm.nih.gov/31319450/)]
38. Gardner J, Warren N. Learning from deep brain stimulation: the fallacy of techno-solutionism and the need for 'regimes of care'. *Med Health Care Philos* 2019 Sep 1;22(3):363-374. [doi: [10.1007/s11019-018-9858-6](https://doi.org/10.1007/s11019-018-9858-6)] [Medline: [30069813](https://pubmed.ncbi.nlm.nih.gov/30069813/)]
39. Longoni C, Bonezzi A, Morewedge CK. Resistance to medical artificial intelligence. *J Consum Res* 2019 Dec;46(4):629-650. [doi: [10.1093/jcr/ucz013](https://doi.org/10.1093/jcr/ucz013)]
40. Berkhout F. Normative expectations in systems innovation. *Technol Anal Strateg Manag* 2006 Jul;18(3-4):299-311. [doi: [10.1080/09537320600777010](https://doi.org/10.1080/09537320600777010)]
41. Konrad K. The social dynamics of expectations: the interaction of collective and actor-specific expectations on electronic commerce and interactive television. *Technol Anal Strateg Manag* 2006 Jul;18(3-4):429-444. [doi: [10.1080/09537320600777192](https://doi.org/10.1080/09537320600777192)]
42. Mittelman M, Markham S, Taylor M. Patient commentary: stop hyping artificial intelligence-patients will always need human doctors. *BMJ* 2018 Nov 07;363:k4669. [doi: [10.1136/bmj.k4669](https://doi.org/10.1136/bmj.k4669)] [Medline: [30404859](https://pubmed.ncbi.nlm.nih.gov/30404859/)]
43. Topol E. *Deep Medicine: How Artificial Intelligence Can Make Healthcare Human Again*. New York, NY: Basic Books; 2019.
44. Verghese A, Shah NH, Harrington RA. What this computer needs is a physician: humanism and artificial intelligence. *JAMA* 2018 Jan 02;319(1):19-20. [doi: [10.1001/jama.2017.19198](https://doi.org/10.1001/jama.2017.19198)] [Medline: [29261830](https://pubmed.ncbi.nlm.nih.gov/29261830/)]
45. Nundy S, Montgomery T, Wachter RM. Promoting trust between patients and physicians in the era of artificial intelligence. *JAMA* 2019 Aug 13;322(6):497-498. [doi: [10.1001/jama.2018.20563](https://doi.org/10.1001/jama.2018.20563)] [Medline: [31305873](https://pubmed.ncbi.nlm.nih.gov/31305873/)]
46. Mittelstadt B. The impact of artificial intelligence on the doctor-patient relationship. Council of Europe. URL: <https://www.coe.int/en/web/bioethics/report-impact-of-ai-on-the-doctor-patient-relationship> [accessed 2023-04-14]
47. Liu X, Keane PA, Denniston AK. Time to regenerate: the doctor in the age of artificial intelligence. *J R Soc Med* 2018 Apr;111(4):113-116 [FREE Full text] [doi: [10.1177/0141076818762648](https://doi.org/10.1177/0141076818762648)] [Medline: [29648509](https://pubmed.ncbi.nlm.nih.gov/29648509/)]
48. Sauerbrei A, Kerasidou A, Lucivero F, Hallowell N. The impact of artificial intelligence on the person-centred, doctor-patient relationship: some problems and solutions. *BMC Med Inform Decis Mak* 2023 Apr 20;23(1):73 [FREE Full text] [doi: [10.1186/s12911-023-02162-y](https://doi.org/10.1186/s12911-023-02162-y)] [Medline: [37081503](https://pubmed.ncbi.nlm.nih.gov/37081503/)]
49. Sparrow R, Hatherley J. High hopes for "deep medicine"? AI, economics, and the future of care. *Hastings Cent Rep* 2020 Jan 18;50(1):14-17. [doi: [10.1002/hast.1079](https://doi.org/10.1002/hast.1079)] [Medline: [32068275](https://pubmed.ncbi.nlm.nih.gov/32068275/)]
50. Sanz Menéndez L, Cabello C. Expectations and learning as principles of shaping the future. Unidad de Políticas Comparadas (CSIC). 2000 Feb. URL: https://digital.csic.es/bitstream/10261/1491/1/expectations_learning.pdf [accessed 2023-04-26]
51. Guice J. Designing the future: the culture of new trends in science and technology. *Res Policy* 1999 Jan;28(1):81-98. [doi: [10.1016/s0048-7333\(98\)00105-x](https://doi.org/10.1016/s0048-7333(98)00105-x)]
52. van Lente H. Navigating foresight in a sea of expectations: lessons from the sociology of expectations. *Technol Anal Strateg Manag* 2012 Sep;24(8):769-782. [doi: [10.1080/09537325.2012.715478](https://doi.org/10.1080/09537325.2012.715478)]

Abbreviations

AI: artificial intelligence

CDSS: clinical decision support system

EXPLaiN: Ethical and Legal issues of Mobile Health-Data: Improving Understanding and Explainability of Digital Transformation and Data Technologies Using Artificial Intelligence

HRA: Human Research Act

Edited by K El Emam, B Malin; submitted 09.06.23; peer-reviewed by H Wu, G Myreteg, L Weidener; comments to author 24.10.23; revised version received 27.01.24; accepted 03.06.24; published 19.08.24.

Please cite as:

Lorenzini G, Arbelaez Ossa L, Milford S, Elger BS, Shaw DM, De Clercq E

The “Magical Theory” of AI in Medicine: Thematic Narrative Analysis

JMIR AI 2024;3:e49795

URL: <https://ai.jmir.org/2024/1/e49795>

doi: [10.2196/49795](https://doi.org/10.2196/49795)

PMID: [39158953](https://pubmed.ncbi.nlm.nih.gov/39158953/)

©Giorgia Lorenzini, Laura Arbelaez Ossa, Stephen Milford, Bernice Simone Elger, David Martin Shaw, Eva De Clercq. Originally published in JMIR AI (<https://ai.jmir.org>), 19.08.2024. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR AI, is properly cited. The complete bibliographic information, a link to the original publication on <https://www.ai.jmir.org/>, as well as this copyright and license information must be included.

Publisher:
JMIR Publications
130 Queens Quay East.
Toronto, ON, M5A 3Y5
Phone: (+1) 416-583-2040
Email: support@jmir.org

<https://www.jmirpublications.com/>