JMIR AI

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

Contents

Tutorial

Original Papers

Viewpoint

Research Letter

Corrigenda and Addenda

Tutorial

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

Chao Yan[1], PhD; Ziqi Zhang[2], PhD; Steve Nyemba[1], MS; Zhuohang Li[2], MS

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

[2]Department 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

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

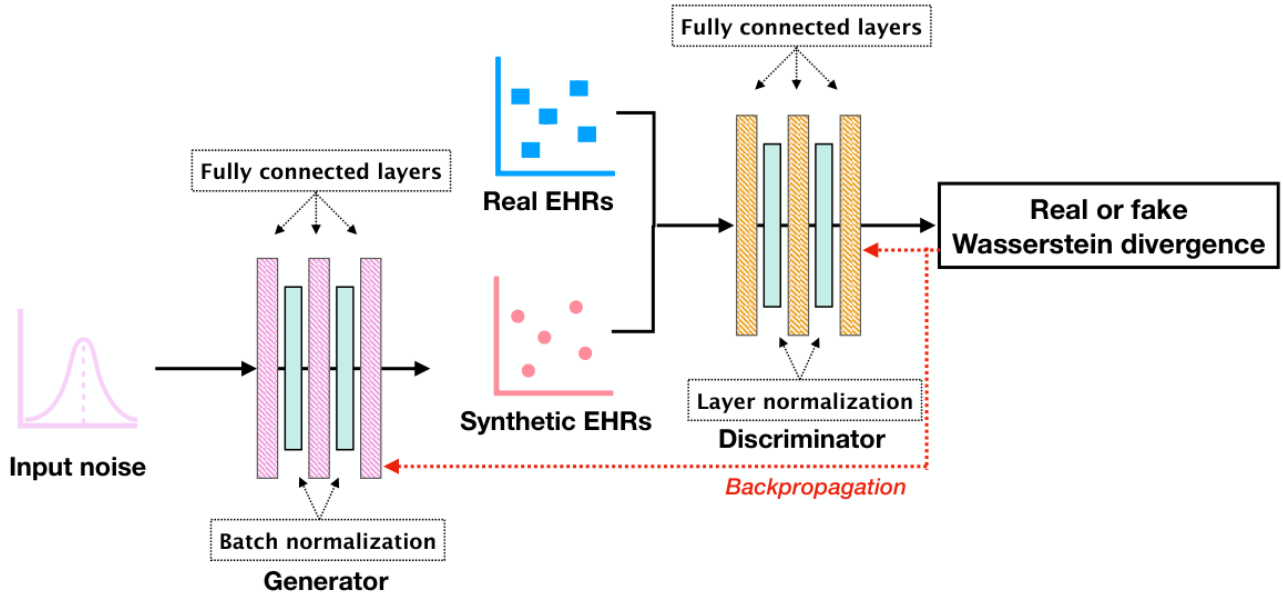
Data Set

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 order among their options, or they may display only a nominal 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.

A

Patient ID	Age	Sex	Race/Ethnicity	Insurance	Died	# of pregnancy	Height (ft)
100001	31	F	White	A	No	1	5.7
100002	68	F	Hispanic	B	Yes	3	5.3
100003	52	F	Asian	C	No	2	5.1

Discrete
Categorical
Continuous
Time (can be converted to continuous variable)

Patient ID	Encounter ID	Timestamp	Diagnosis
100001	201	2018-10-02	Type 2 diabetes
100001	202	2019-05-19	Hypertension
100002	301	2022-12-21	Dementia

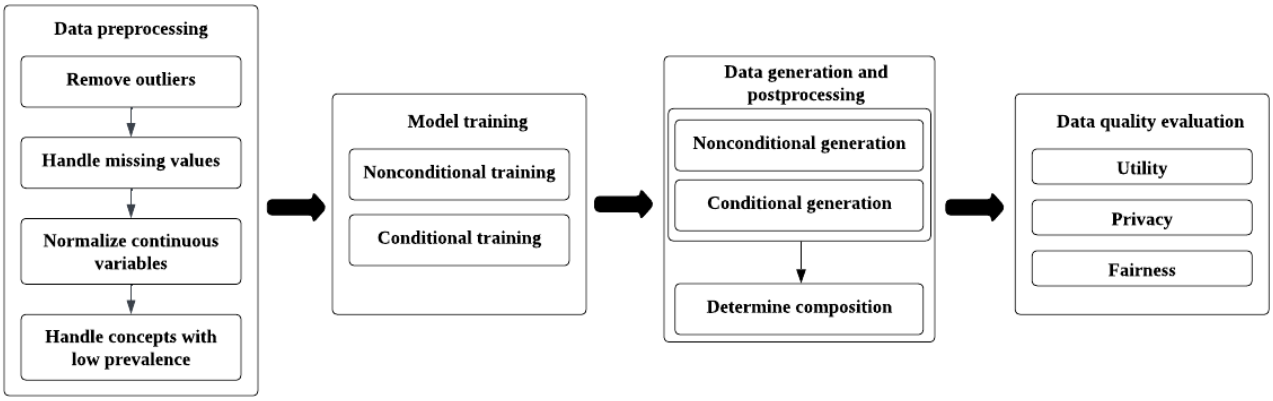
Patient ID	Encounter ID	Timestamp	Medication
100001	201	2018-10-02	Insulin
100001	202	2019-05-19	Acebutolol
100002	301	2022-12-21	Donepezil

Patient ID	Encounter ID	Timestamp	Measurement	Result
100001	201	2018-10-02	Hemoglobin A1C	5.9 %
100001	202	2019-05-19	Blood pressure	145/90 mmHg
100002	301	2022-12-21	Albumin	5.1 g/dL

B

Patient ID	Age	Sex	White	Hispanic	Asian	Ins A	Ins B	Ins C	Died	#P =1	#P =2	#P =3	Height	T2D	HT	DM	Insulin	Acebutolol	Donepezil	H-A1C	BP-S	BP-D	Albumin
100001	31	1	1	0	0	1	0	0	1	1	0	0	5.7	1	1	0	1	1	0	5.9	145	90	NA
100002	68	1	0	1	0	0	1	0	0	0	0	1	5.3	0	0	1	0	0	1	NA	NA	NA	5.1
100003	52	1	0	0	1	0	0	1	1	0	1	0	5.1	0	0	0	0	0	0	NA	NA	NA	NA

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 (277.03)	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 lipoid 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.

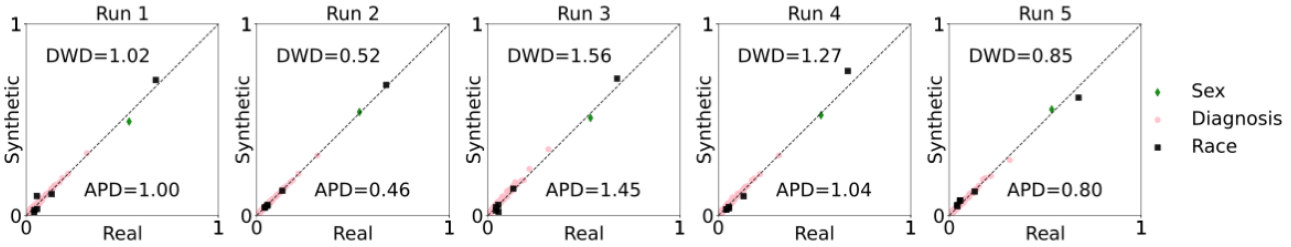
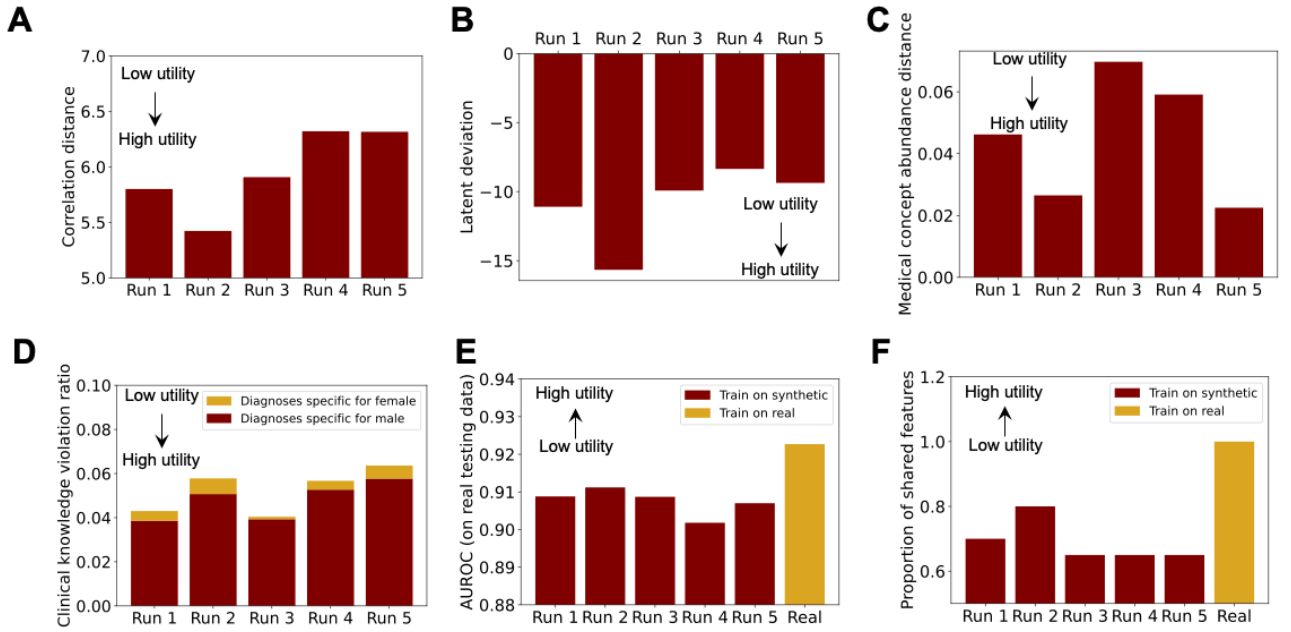


Figure 5. Data utility in (A) column-wise correlation, (B) latent cluster analysis, (C) medical concept abundance, (D) clinical knowledge violation, (E) prediction performance, and (F) feature importance. For clinical knowledge violation, “hyperplasia of prostate," “cancer of prostate,” and “erectile 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

[a]N/A: not applicable.

[b]Indicates 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 evolvement 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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]
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] [Medline: 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]
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] [Medline: 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] [Medline: 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] [Medline: 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]
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]
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]
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]
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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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]
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]
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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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]
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]
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] [Medline: 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] [Medline: 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]
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] [Medline: 34573790]

41. Hashemi AS, Soliman A, Lundström J, Etminani K. Domain knowledge-driven generation of synthetic healthcare data. Stud Health Technol Inform 2023;302:352-353. [doi: 10.3233/SHTI230136] [Medline: 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] [Medline: 31812132]
43. Hashemi AS, Etminani 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]

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
PMID:38875595

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

[1]University of Virginia, Charlottesville, VA, United States

[2]Division of Infectious Diseases & International Health, Department of Medicine, University of Virginia School of Medicine, Charlottesville, VA, United States

[3]Department of Emergency Medicine, Johns Hopkins School of Medicine, Baltimore, MD, United States

[4]Center for Disease Dynamics, Economics and Policy, Washington, DC, DC, United States

[5]Department of Statistics, Virginia Tech, Blacksburg, VA, United States

[6]Department 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.

(JMIR AI 2024;3:e48067) doi:10.2196/48067

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, $S_{p,t}(\alpha) = \{p': (u_p, u_{p'}) \in E, p' \text{ is labeled positive at time } t' \in \{t - a, t]\}$, 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, $\S_{p,t}(\beta) = \{h: (u_p, u_h) \in E, h \text{ visited } p \text{ at time } t' \in (t - \beta, t]\}$. We refer to $|\S_{p,t}(\beta)|$ as Provider β.
- MRSA positive patients collocated 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 $f_{p,t}(u, \gamma) = \{p':(u_p, u_{p'}) \in E, p'$ is labeled positive at $t' \in (t' - \gamma, t]$ and is in the hospital unit u with $p\}$. We referred to $|f_{p,t}(u,\gamma)|$ as the number of patients with colocated MRSA.
- *Bed reuse* Π: let $\Pi_{p,t}(x) = \{p': (u_p, u_{p'}) \notin E, p'$ is labeled positive at time $t'<t$ 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–t1. For the d-days ($d \geq 0$) ahead model, we computed the corresponding length of stay (before the MRSA test) as $\max\{t\text{-}d\text{-}t_1, 0\}$. Note that $t\text{-}d\text{-}t_1$ 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 $e_\delta(t)=\{\exists_{t1}, t1 \in T(e), t_1<t, 0\leq(t-t_1)\leq\delta\}$. We considered δ=90 and e∈ {Surgery, Device implant, Antibiotic, Kidney dialysis}. For the d-days ahead model, the feature is defined by considering δ+d as the parameter in the aforementioned definition, instead of δ.

5. Department-based features: we constructed the following features associated with room stays:

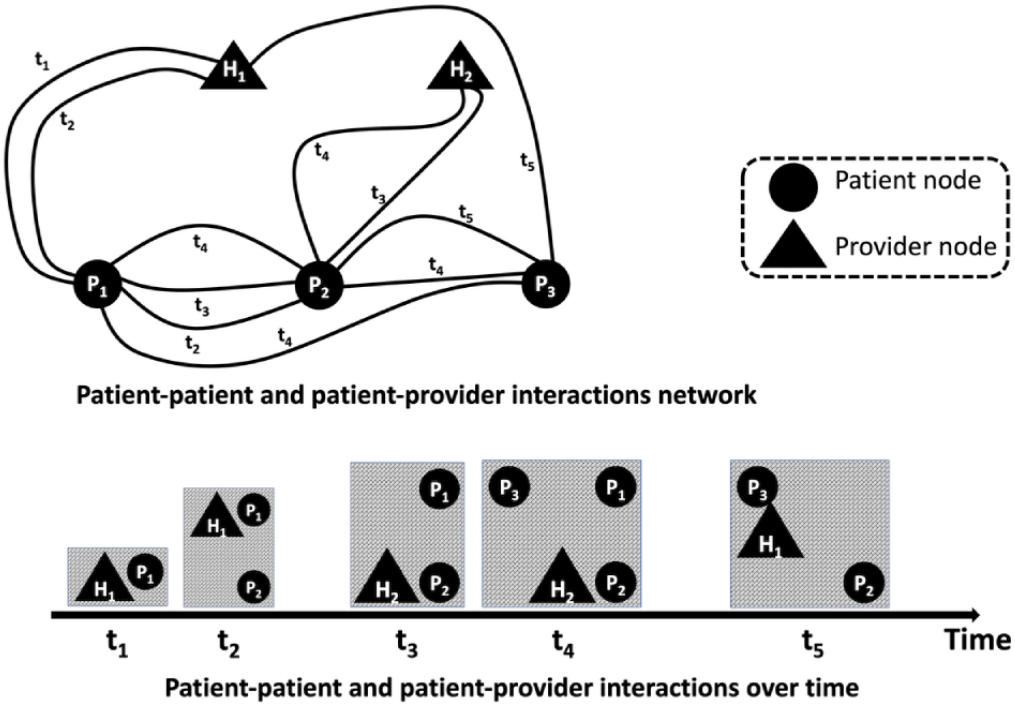
- ICU: this is a Boolean value that indicates whether a patient is admitted to an ICU.
- Emergency department (ED): this is a Boolean value that indicates whether a patient is admitted to the ED.

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.

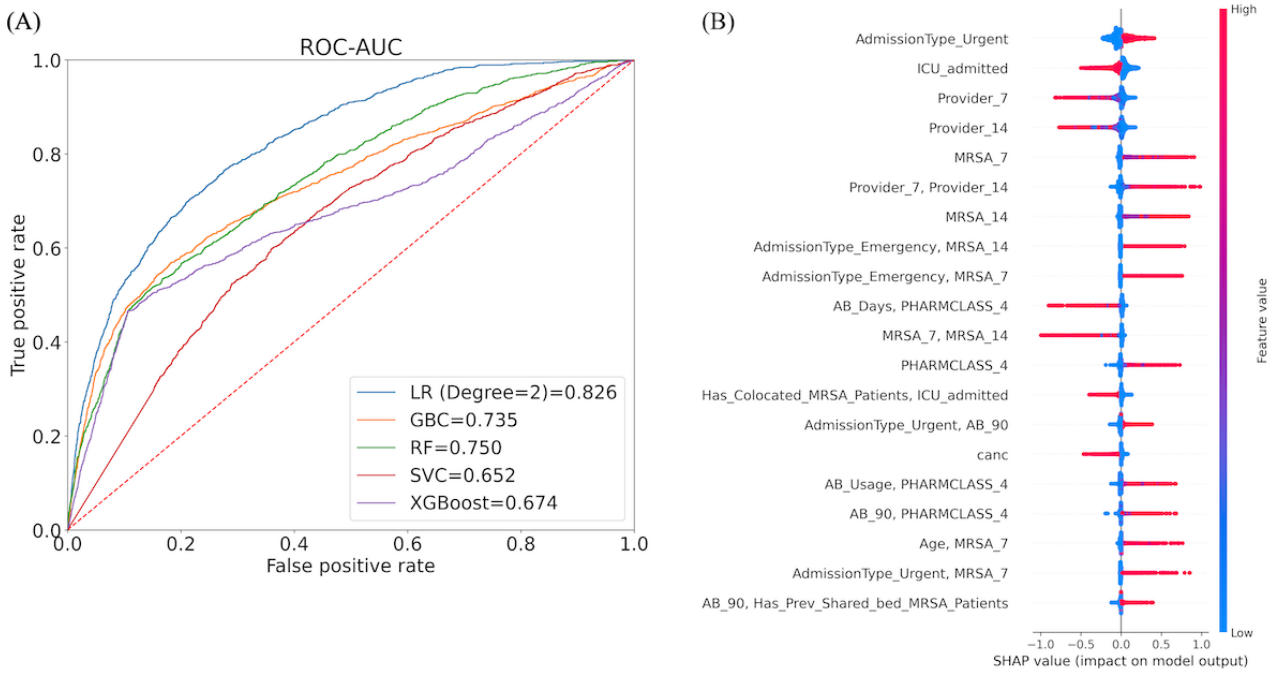
Model Explainability Using SHAP

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 $H^{j}_{p,t}=\{K_{p,ti}: 1 \le i \le j,\ t_j<t_{j-1}<...<t_1<t\}$. No testing history exists for a newly admitted patient, expressed as $H_{p,t}=ø$. The testing history, considering only the last test result, is expressed as $H^{1}_{p,t}=\{K_{p,t1}\}$. Similarly, the testing history, considering the last 2 test results, is expressed as $H^{2}_{p,t}=\{K_{p,t2}\}$. 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

[a]ICU: intensive care unit.

[b]ED: 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 ≥ 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

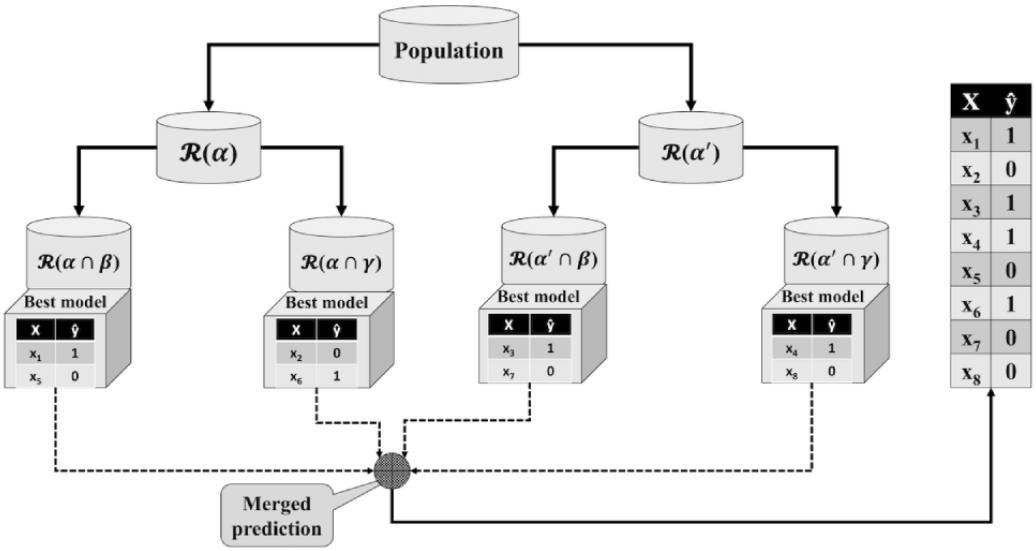
Figure 3 shows the schematic architecture of the hierarchical model. The construction steps of the hierarchical model are as follows:

- 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(α)=patient subgroup of 0 to 50 years old and (2) R(α')=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: $P(\alpha) \cap P(\alpha') = \varnothing$
  - 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(β)=patient subgroup of ICU and (2) R(γ)=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, Xi 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(α∩β) 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 $d \in \{1,2,\ldots,7\}$, 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.

3. 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).
4. "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	*0.936* [j]	*0.882*	*0.878*	*0.886*	*0.800*	*0.114*	0.067	*0.837*	*0.749*
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	*0.034*	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	*0.883*	0.490	*0.807*	0.832	0.440	0.168	*0.037*	0.569	0.507

[a]This column specifies the best-performing model.

[b]ROC-AUC: receiver operating characteristics-area under the curve.

[c]AUPRC: area under the precision-recall curve.

[d]FPR: false positive rate.

[e]FNR: false negative rate.

[f]MCC: Matthews correlation coefficient.

[g]LR: penalized logistic regression.

[h]ICU: intensive care unit.

[i]ED: emergency department.

[j]The best value for each performance metric is italicized.

[k]HCF: health care facility.

[l]For "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

[a]The "Model" column specifies the best-performing model (LR=penalized logistic regression classifier, GB=gradient boosting, and SV=support vector).

[b]ROC-AUC: receiver operating characteristics-area under the curve.

[c]AUPRC: area under the precision-recall curve.

[d]FPR: false positive rate.

[e]FNR: false negative rate.

[f]MCC: Matthews correlation coefficient.

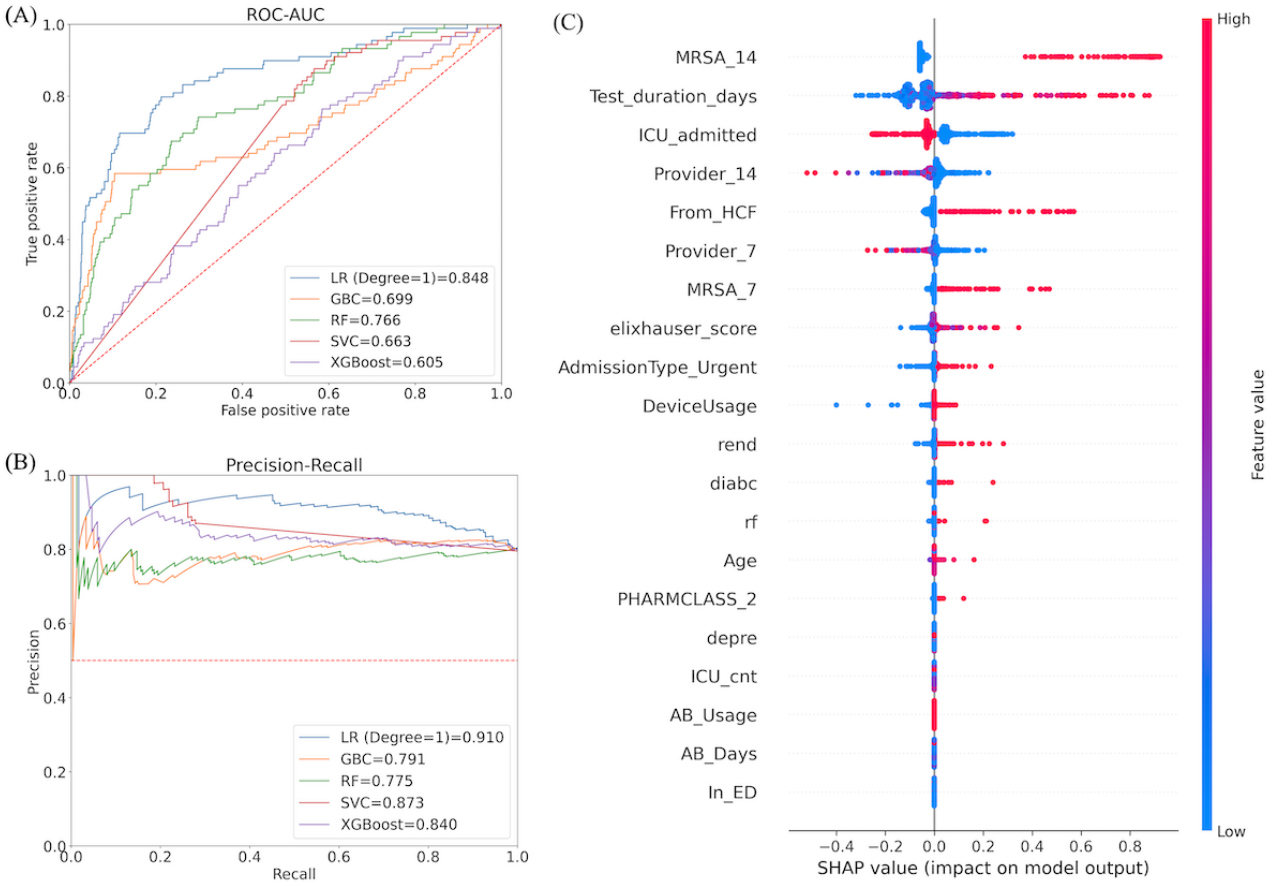
[g]LR: logistic regression.

[h]GB: gradient boosting.

[i]The best value for each performance metric is italicized.

[j]SV: 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.

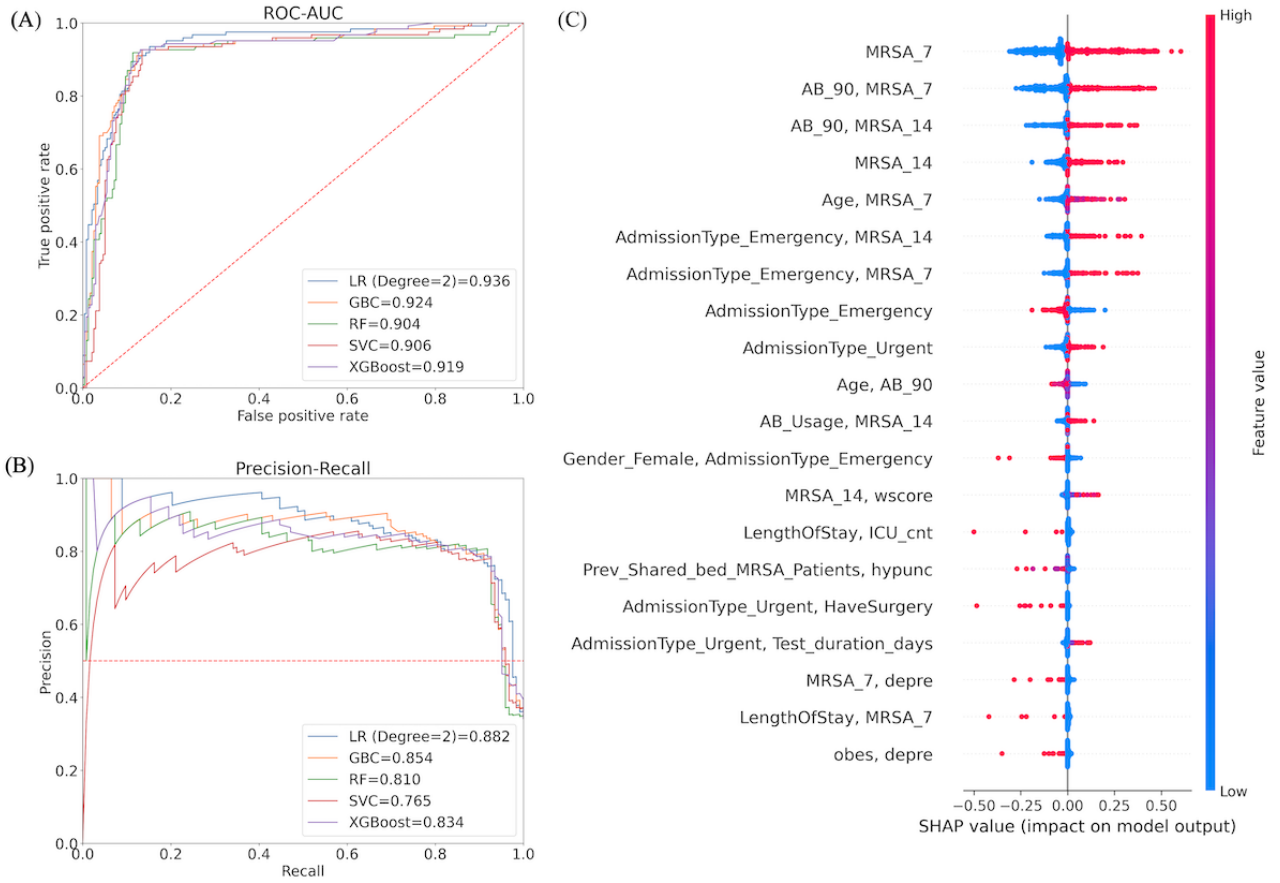
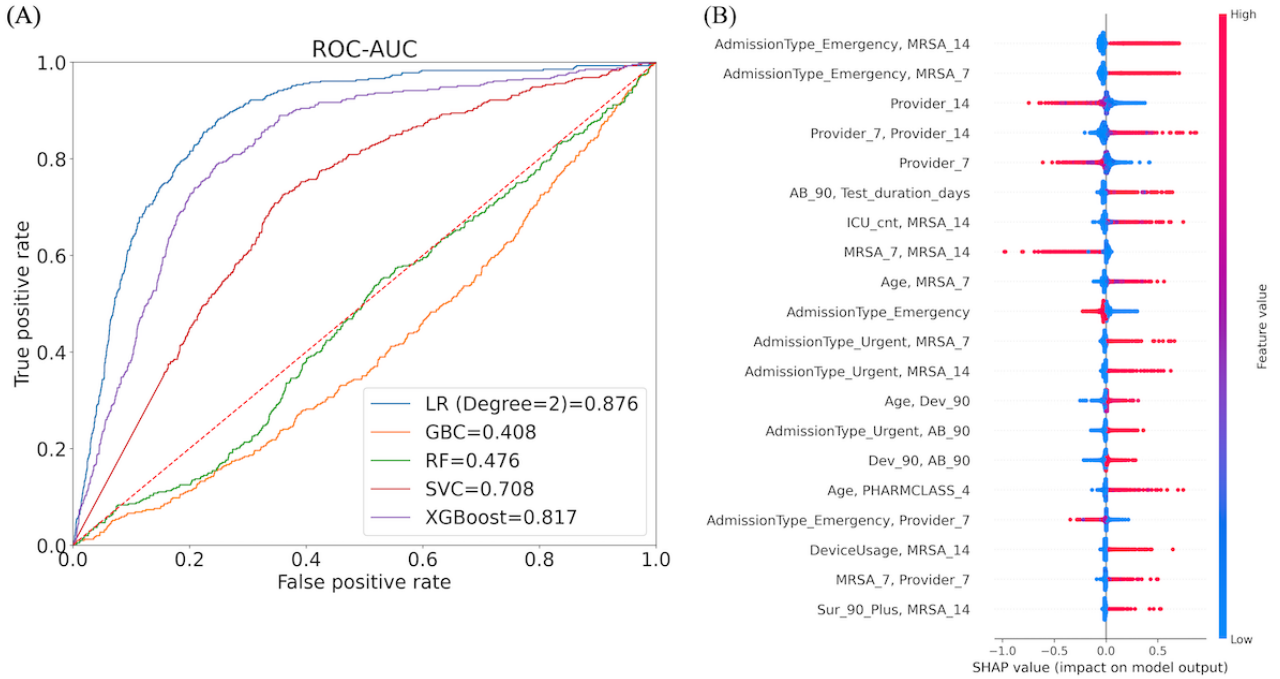


Figure 6. (A) Performance of different machine learning models for the intensive care unit subpopulation; the penalized logistic regression (LR) model performs best. (B) Top features of the LR model. GBC: gradient boosted classifier; RF: random forest; SHAP: Shapley Additive Explanations; SVC: support 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

[a]The "Model" column specifies the best-performing model (LR=penalized logistic regression classifier).

[b]AUC: area under the curve.

[c]AUPRC: area under the precision-recall curve.

[d]FNR: false negative rate.

[e]MCC: Matthews correlation coefficient.

[f]LR: logistic regression.

[g]ICU: intensive care unit.

[h]ED: emergency department.

[i]HCF: health care facility.

[j]italics.

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

[a]The "Model" column specifies the best-performing model (LR=penalized logistic regression, GB=gradient boosting, and SV=support vector).

[b]AUC: area under the curve.

[c]AUPRC: area under the precision-recall curve.

[d]FNR: false negative rate.

[e]MCC: Matthews correlation coefficient.

[f]LR: logistic regression.

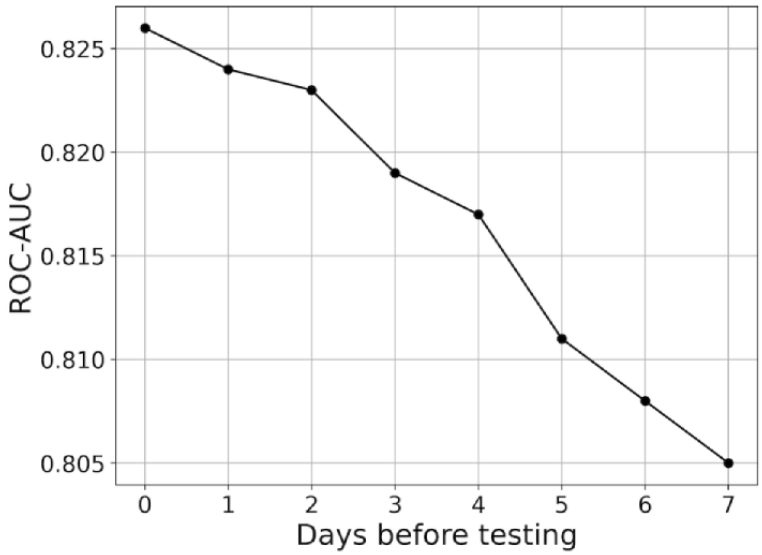
[g]GB: gradient boosting.

[h]SV: 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 hern 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,antibiotic%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]
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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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]
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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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]
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] [Medline: 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
PMID:38875598

Original Paper

Framework for Ranking Machine Learning Predictions of Limited, Multimodal, and Longitudinal Behavioral Passive Sensing Data: Combining User-Agnostic and Personalized Modeling

Tahsin Mullick[1], MEng; Sam Shaaban[2], MBA; Ana Radovic[3], MD, MSc; Afsaneh Doryab[1], PhD

[1]Department of Systems and Information Engineering, University of Virginia, Charlottesville, VA, United States

[2]NuRelm, Pittsburgh, PA, United States

[3]Department 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

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

[a]SOTA: state of the art.
[b]ML: machine learning.
[c]CV: cross-validation.
[d]HAR: human activity recognition.
[e]DT: decision tree.
[f]LR: linear regression.
[g]MLP: multilayer perceptron.
[h]KNN: K-nearest neighbor.
[i]SVM: support vector machine.
[j]RF: random forest.
[k]XGBoost: Extreme Gradient Boosting.
[l]NB: 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.

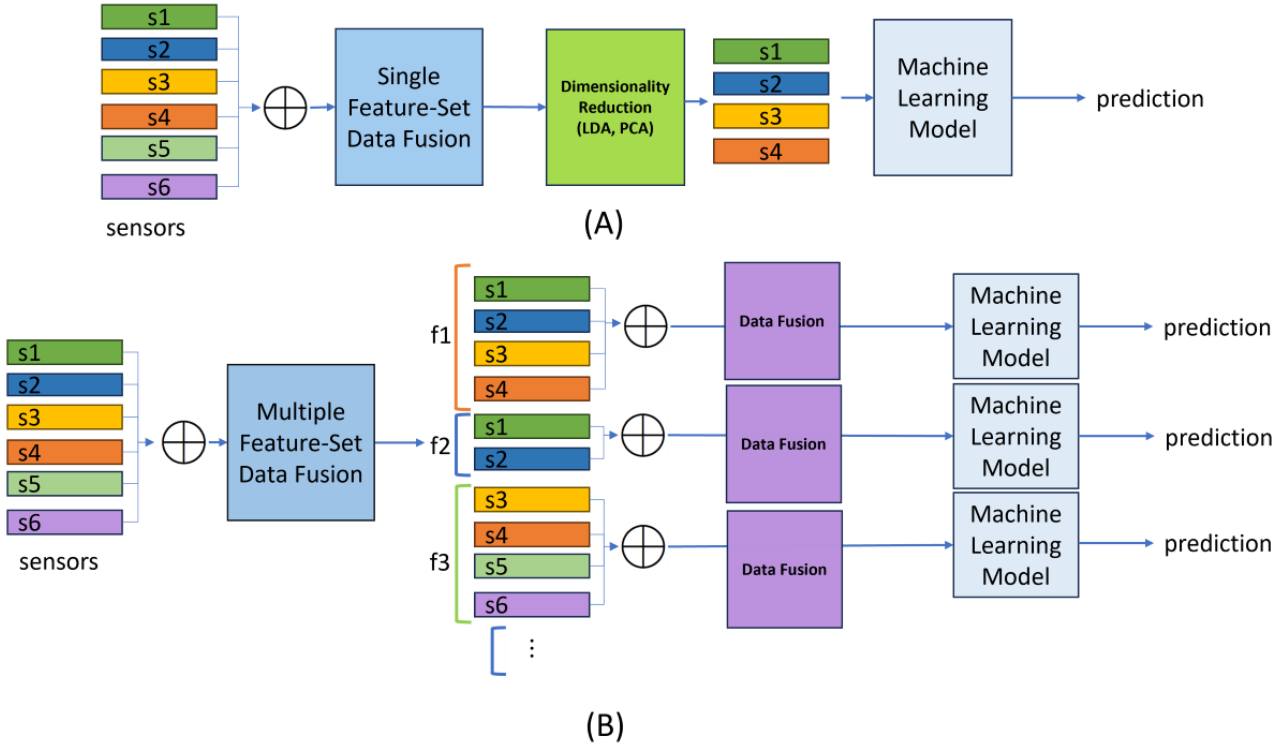
Role of Data Fusion in Passive Sensing Studies

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

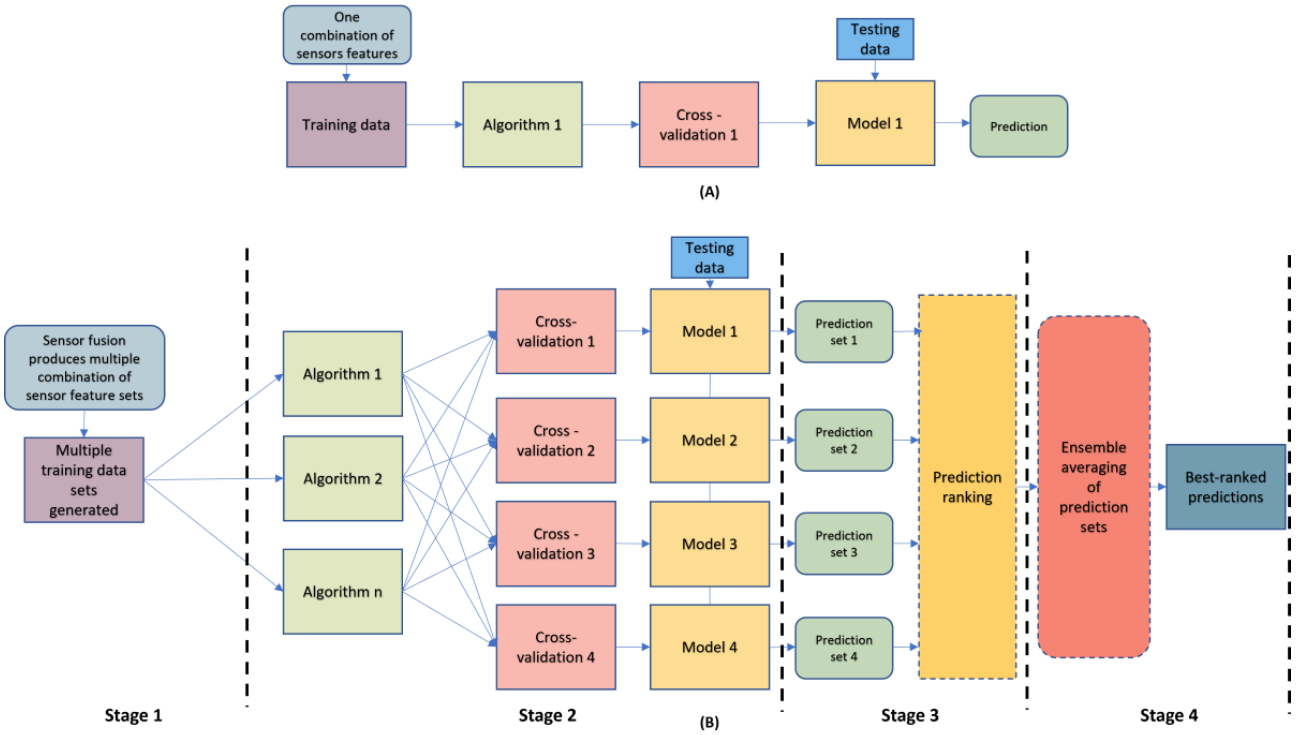
We proposed a framework for longitudinal multimodal sensors (FLMS) as a ranking framework to rigorously handle longitudinal, multimodal sensor data and incorporate different analysis and modeling strategies suited for small and sparse time series data sets to produce better results. The FLMS incorporates 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 $\sum_{x=1}^{Z} \binom{Z}{x}$, 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 $\sum_{x=1}^{6} \binom{6}{x}$.

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:

(1)

,where alg refers to the algorithm with the lowest absolute sum error and $pred_m(alg_t)$ 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.

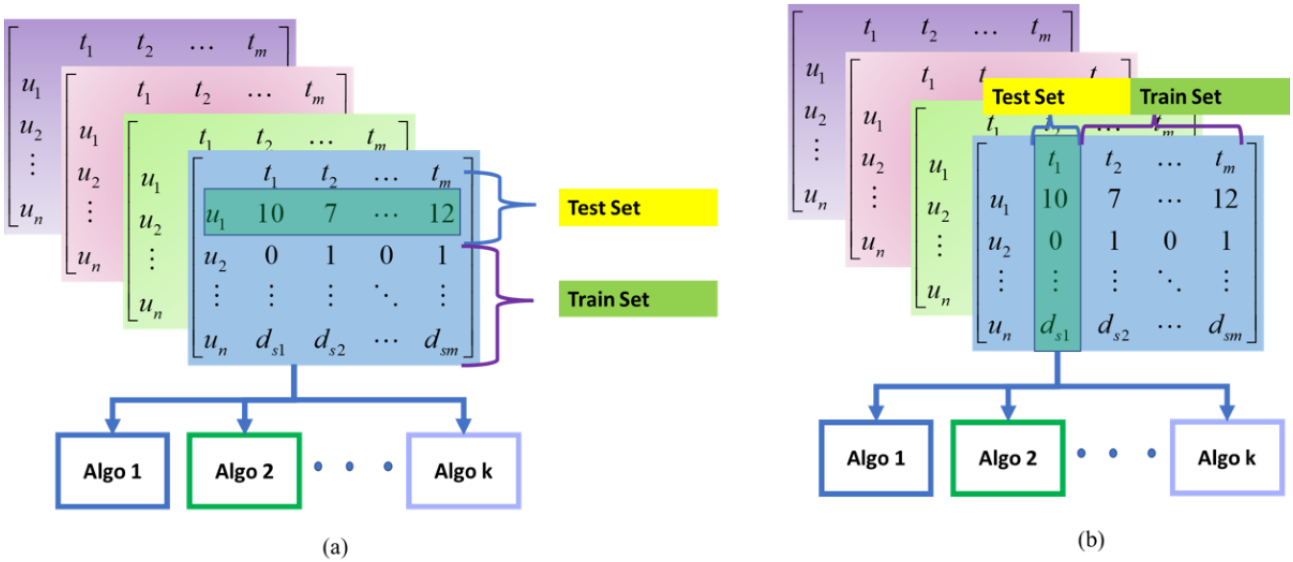
User-Agnostic Model Building

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

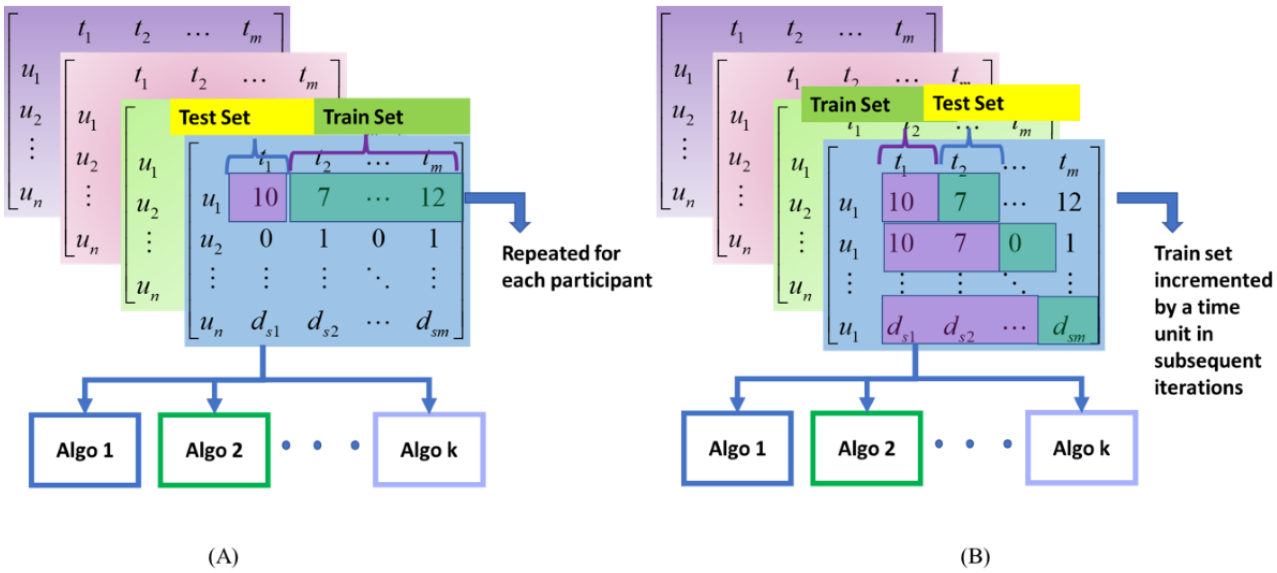
The personalized modeling strategy leverages each user's historical and cross-time data samples in a sliding window and the leave-one-time-unit-out approach.

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:

(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  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:

(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.

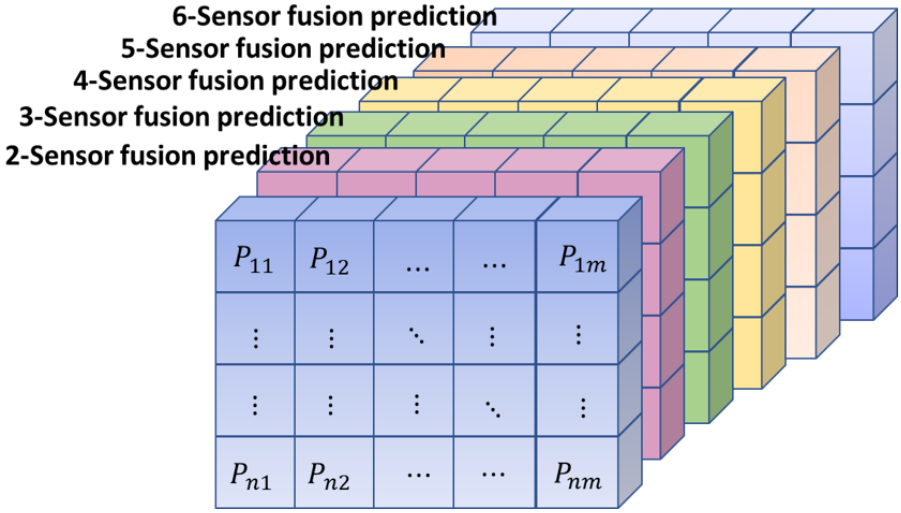


Figure 6. Instance of ATU where it shows how the mean aggregated prediction set is generated according to Equation (2). ATU: accumulated time unit; avg: average.

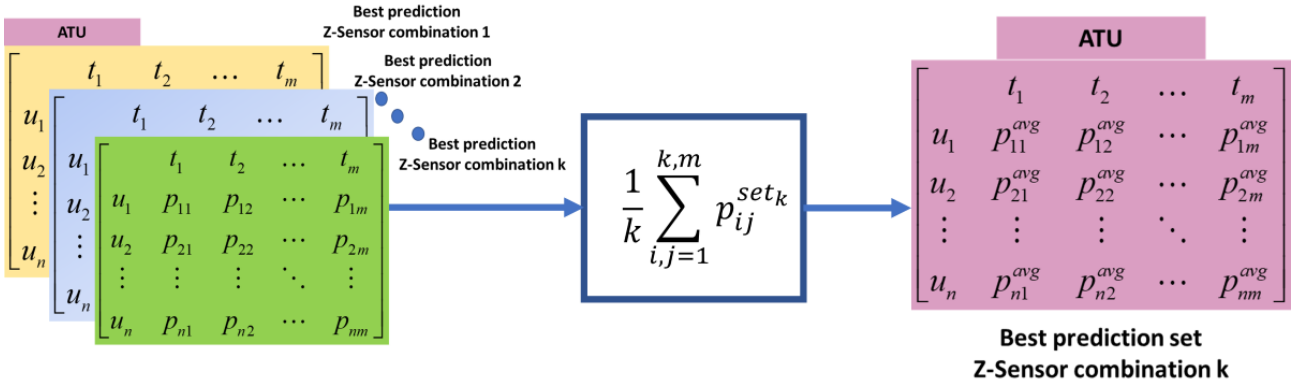


Figure 7. The 0-1 encoding process resolves dealing with large variances in regression values. ATU: accumulated time unit; LOPO: leave one participant out; LOTPO: leave one time unit of participant out; LTXO: leave time unit X out.

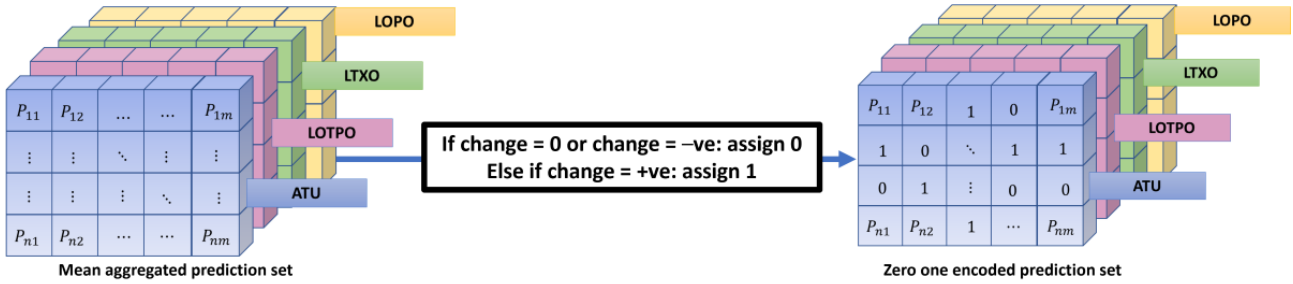
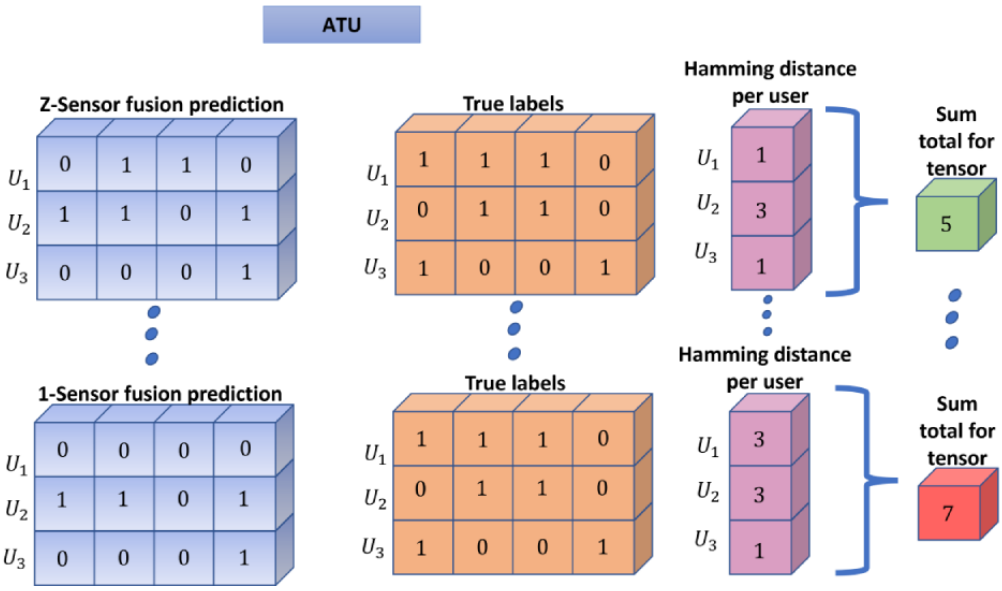


Figure 8. Hamming distance calculations reduce error and also determine the weight each of the 4 modeling approaches will contribute to stage 4's ensembled weighted average. ATU: accumulated time unit.



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):

(4)

,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.

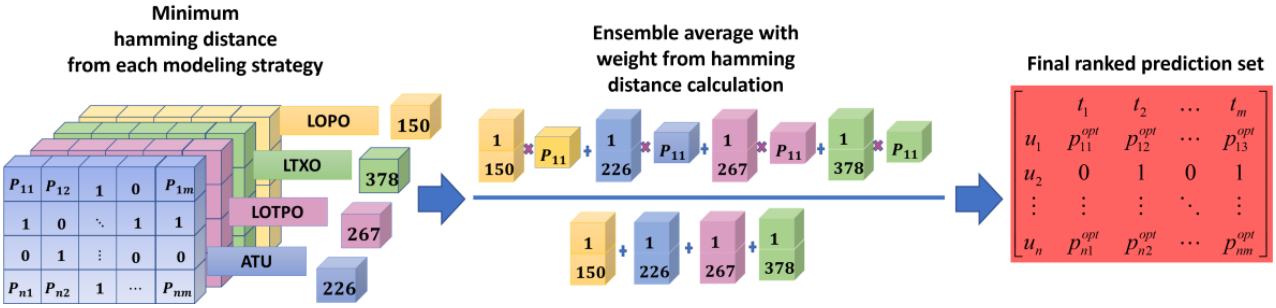
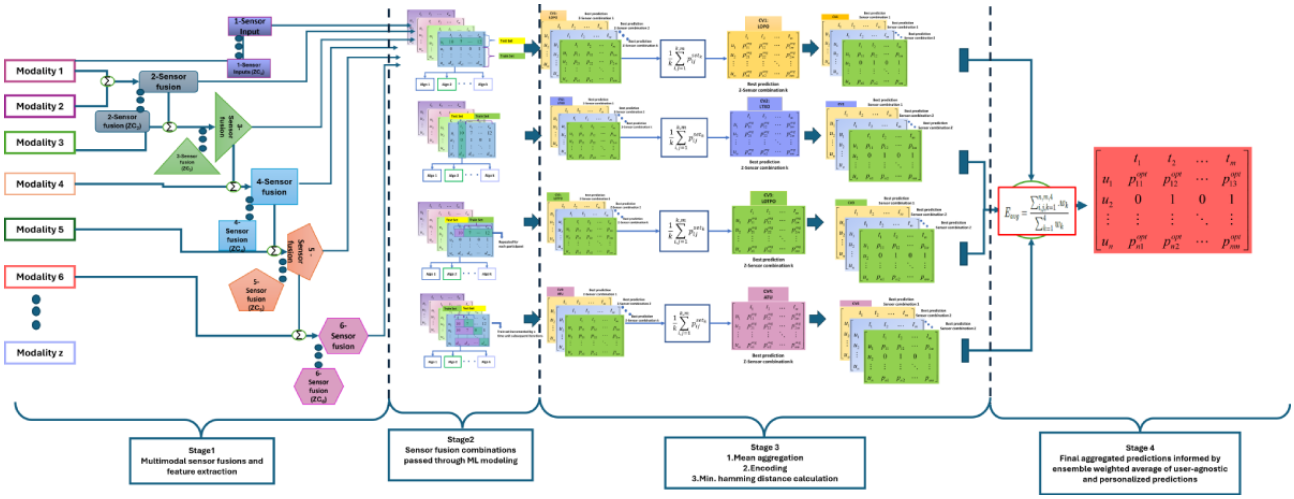


Figure 10. FLMS ranking overview. Algo: algorithm; ATU: accumulated time unit; avg: average; CV: cross-validation; FLMS: framework for longitudinal multimodal sensors; LOPO: leave one participant out; LOTPO: leave one time unit of participant out; LTXO: leave time unit X out; ML: machine learning.



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 corelated 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)

[a]FLMS: framework for longitudinal multimodal sensors.

[b]ATU: accumulated time unit.

[c]LOTPO: leave one time unit one participant out.

[d]LOPO: leave one participant out.

[e]LTXO: 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

[a]FLMS: framework for longitudinal multimodal sensors.

[b]ATU: accumulated time unit.

[c]LOTPO: leave one time unit one participant out.

[d]LOPO: leave one participant out.

[e]LTXO: 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

[a]KNN: K-nearest neighbor.

[b]CV: cross-validation.

[c]FLMS: framework for longitudinal multimodal sensors.

[d]ATU: accumulated time unit.

[e]LOTPO: leave one time unit one participant out.

[f]LTXO: leave time unit X out.

[g]LOPO: leave one participant out.

[h]ML: machine learning.

[i]XGBoost: 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.

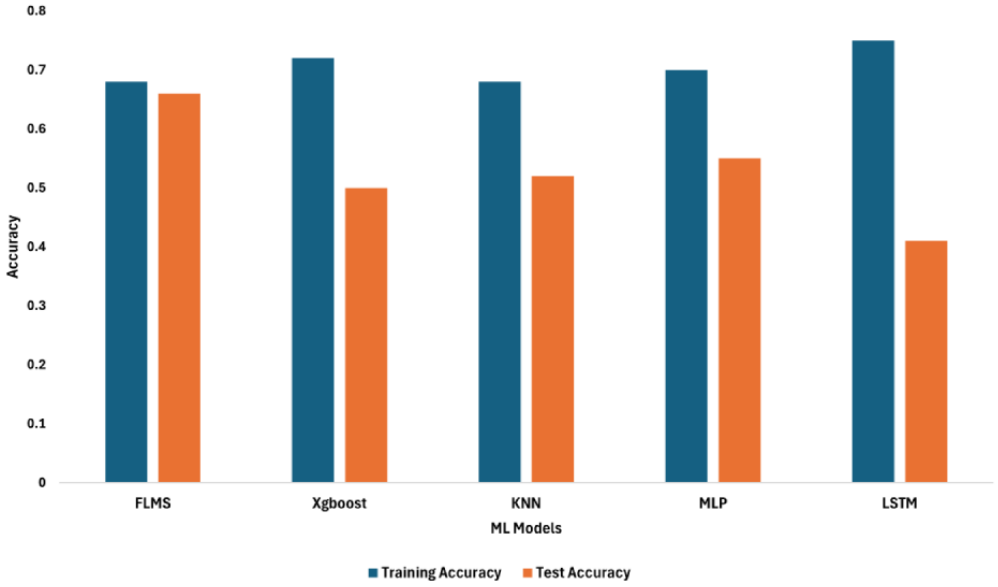


Table 5. Experiment 2 performance of the FLMS[a] compared to ML[b] and DL[c] algorithms used in the current literature on adolescent data.

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)

[a]FLMS: framework for longitudinal multimodal sensors.
[b]ML: machine learning.
[c]DL: deep learning.
[d]ATU: accumulated time unit.
[e]LOTPO: leave one time unit one participant out.
[f]LOPO: leave one participant out.
[g]LTXO: leave time unit X out.
[h]XGBoost: Extreme Gradient Boosting.
[i]KNN: K-nearest neighbor.
[j]MLP: multilayer perceptron.
[k]LSTM: 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

[a]KNN: K-nearest neighbor.

[b]CV: cross-validation.

[c]FLMS: framework for longitudinal multimodal sensors.

[d]ML: machine learning.

[e]XGBoost: Extreme Gradient Boosting.

[f]LSTM: 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)

[a]FLMS: framework for longitudinal multimodal sensors.
[b]ML: machine learning.
[c]ATU: accumulated time unit.
[d]LOTPO: leave one time unit one participant out.
[e]LOPO: leave one participant out.
[f]LTXO: leave time unit X out.
[g]XGBoost: Extreme Gradient Boosting.
[h]KNN: K-nearest neighbor.
[i]MLP: multilayer perceptron.
[j]SVM: support vector machine.
[k]DT: decision tree.
[l]RF: random forest.
[m]AdaBoost: Adaptive Boosting.

Table 8. Experiment 3 metadata.

Metadata	Experiment 3
Feature preprocessing	KNN[a] imputation, dropping highly corelated 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

[a]KNN: K-nearest neighbor.
[b]CV: cross-validation.
[c]FLMS: framework for longitudinal multimodal sensors.
[d]ML: machine learning.
[e]XGBoost: Extreme Gradient Boosting.
[f]SVM: 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 1K23MH111922-01A1). 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] [Medline: 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]
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]
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]
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]
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]
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]
8. Dietterich T. Overfitting and undercomputing in machine learning. ACM Comput Surv 1995 Sep;27(3):326-327. [doi: 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]
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]
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]
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]
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]

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]
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]
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]
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] [Medline: 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]
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] [Medline: 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]
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]
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]
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]
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]
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]
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]
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]
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] 2021 [FREE Full text] [doi: 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]
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] [Medline: 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]
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] [Medline: 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] [Medline: 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]

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]
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] [Medline: 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]
40. Pedrelli P, Fedor S, Ghandeharioun A, Howe E, Ionescu DF, Bhathena 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/fpsyt.2020.584711] [Medline: 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]
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] [Medline: 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]
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] [Medline: 34610915]
45. Vega J, Li M, Aguillera 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/fdgth.2021.769823] [Medline: 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] [Medline: 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]
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] [Medline: 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]
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]
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]
52. Lane ND, Lin M, Mohammod 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]
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]
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] [Medline: 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]

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]
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]
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]
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]
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]
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] 2021. [doi: 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] [Medline: 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] [Medline: 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/automatically-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]
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]
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]
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] 2021;07985(2021). [doi: 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] 2021. [doi: 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
PMID:38875667

Original Paper

Sepsis Prediction at Emergency Department Triage Using Natural Language Processing: Retrospective Cohort Study

Felix Brann[1], BSc; Nicholas William Sterling[1], MD, PhD, MS; Stephanie O Frisch[1], PhD, RN; Justin D Schrager[1,2], MD, MPH

[1]Vital Software, Inc, Claymont, DE, United States

[2]Department 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 where 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

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)

[a]Sepsis 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.

[b]BP: blood pressure.

[c]SpO_2: oxygen saturation.

[d]WBC: 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.

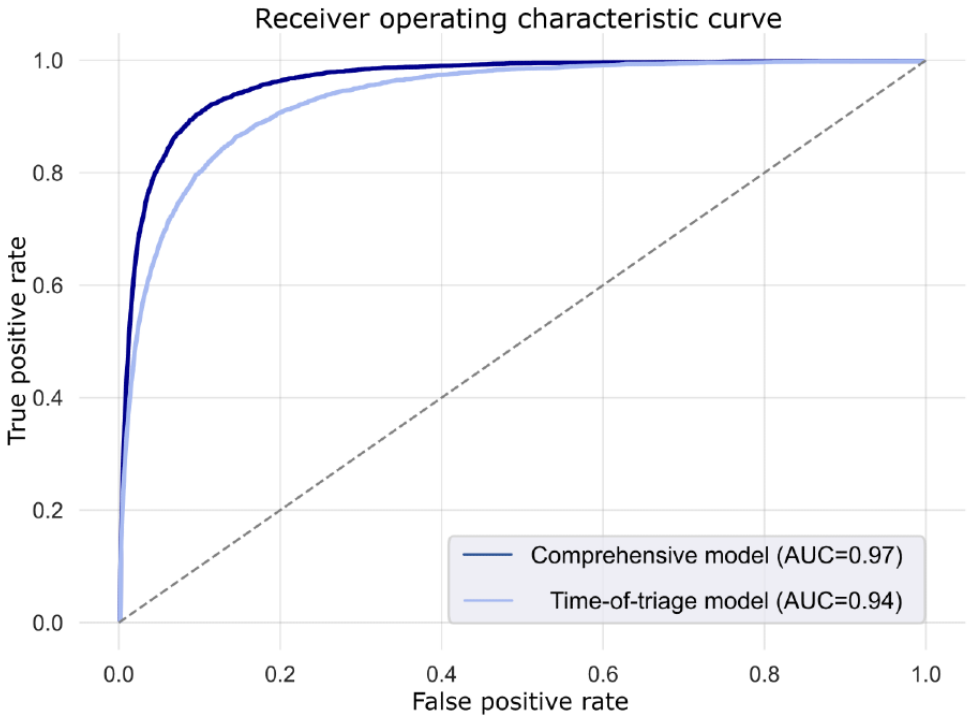
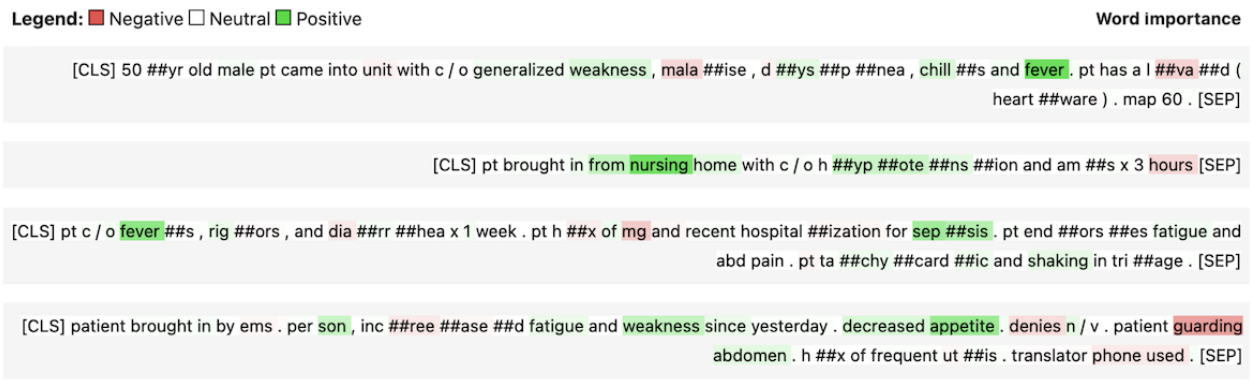


Figure 2. Heat maps applied to words and subwords of a sample of emergency department nursing triage notes to indicate relative contributions to sepsis prediction.



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.

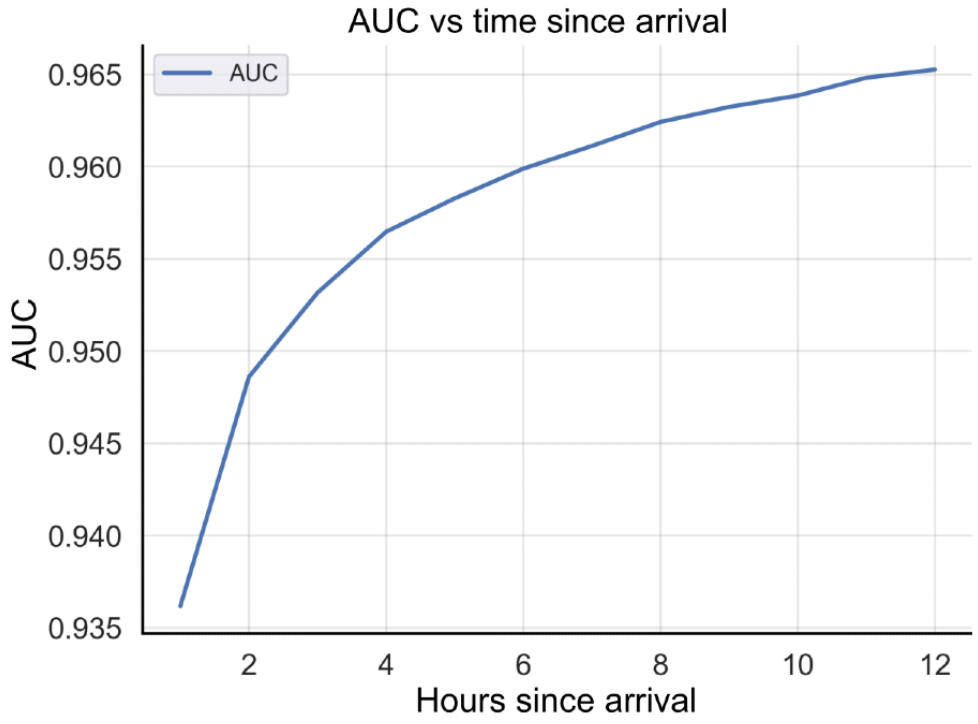


Table 2. Machine learning prediction of sepsis using data available at the time of emergency department (ED) triage ("time-of-triage" model) and all data available after ED workup ("comprehensive" model).

	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

[a]AUC: area under the receiver operating characteristic curve.

[b]CDC: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 33230117]
12. Wong A, Otles 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] [Medline: 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] [Medline: 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] [Medline: 33392576]
17. Sterling NW, Patzer RE, Di M, Schrager 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] [Medline: 31445253]
18. Zhang X, Kim J, Patzer RE, Pitts SR, Patzer A, Schrager 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] [Medline: 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] [Medline: 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]
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] [Medline: 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, Schrager 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
PMID:38875594

Original Paper

Understanding the Long Haulers of COVID-19: Mixed Methods Analysis of YouTube Content

Alexis Jordan[1], MS; Albert Park[1], 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

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 conspiracytheories*, *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*, *safetyconcerns*, *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*, *diagnosticconcerns*, *broadhealthconcerns*, *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 destigmatize 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," "heartbeat," "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 haywire 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 thc 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."

[a]CBD: cannabidiol.

[b]THC: 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 Biterm 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 hyperestesia?" 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] [Medline: 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] [Medline: 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]
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] [Medline: 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] [Medline: 35412189]
8. Oyebode 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] [Medline: 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] [Medline: 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]
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] [Medline: 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] [Medline: 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]
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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 34860328]
18. Uddin SM, Albert A, Tamanna M, Alsharef 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]
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] [Medline: 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]
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] [Medline: 34762060]
25. Yan X, Guo J, Lan Y, Cheng X. A biterm 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]
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]
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]
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] [Medline: 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]
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] [Medline: 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]
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] [Medline: 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] [Medline: 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] [Medline: 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]
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] [Medline: 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] [Medline: 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]
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] [Medline: 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] [Medline: 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
PMID:38875666

Original Paper

Identifying Links Between Productivity and Biobehavioral Rhythms Modeled From Multimodal Sensor Streams: Exploratory Quantitative Study

Runze Yan[1], PhD; Xinwen Liu[2], MS; Janine M Dutcher[2], PhD; Michael J Tumminia[3], PhD; Daniella Villalba[2], PhD; Sheldon Cohen[2], PhD; John D Creswell[2], PhD; Kasey Creswell[2], PhD; Jennifer Mankoff[4], PhD; Anind K Dey[4], PhD; Afsaneh Doryab[1], PhD

[1]University of Virginia, Charlottesville, VA, United States
[2]Carnegie Mellon University, Pittsburgh, PA, United States
[3]University of Pittsburgh, Pittsburgh, PA, United States
[4]University 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

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.

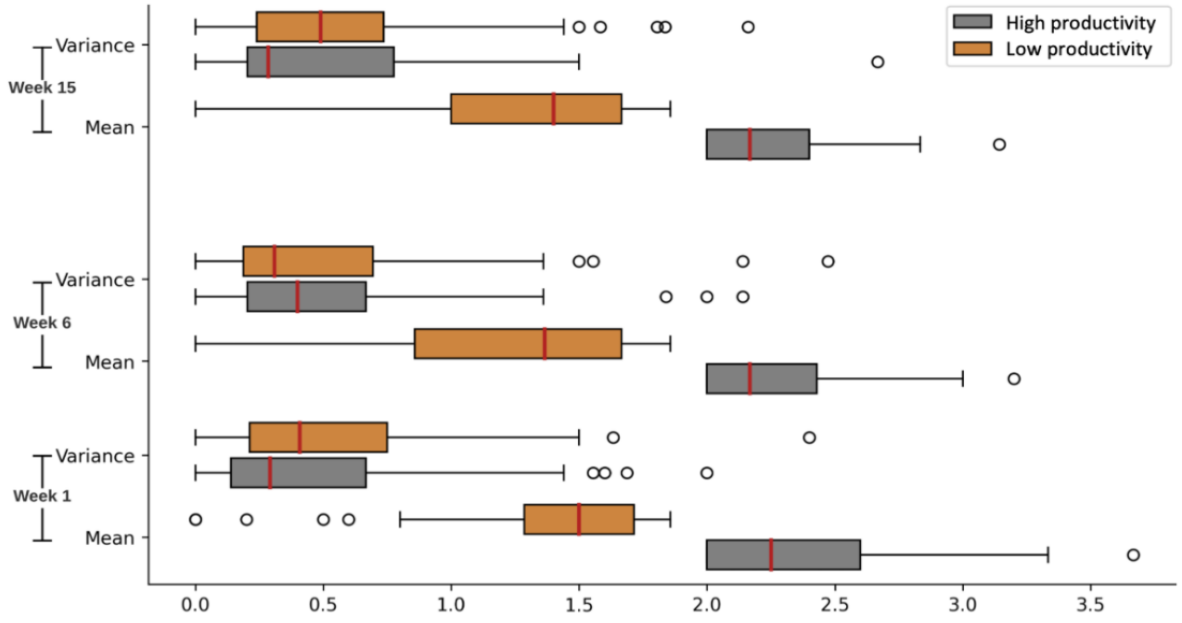


Table 2. Participant productivity[a].

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

[a]The 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 Biobehavioral Rhythms

To model rhythms from longitudinal biobehavioral 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; is the Midline Estimating Statistic of Rhythm (MESOR); t_i is the sampling time; is the number of input periods; A_c, T_c, and $Ø_c$ represent the amplitude (Amp), period, and acrophase (PHI), respectively; and is the error. Cosinor also outputs the SE for MESOR, Amp, and PHI, respectively.

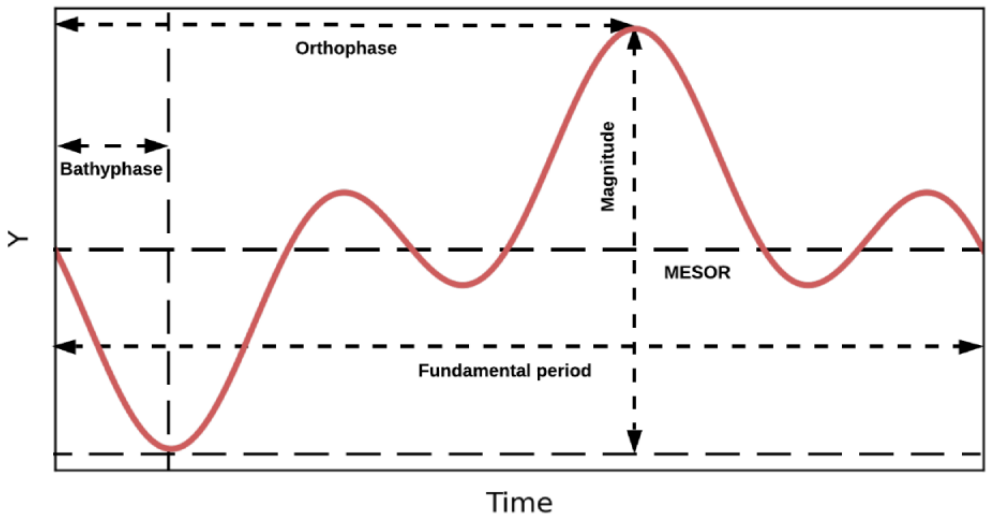
We used Cosinor to build personal cyclic models per student per sensor stream in weeks 1, 6, 15, and the weeks adjacent to them (eg, for week 6, we use sensor data from weeks 5, 6, and 7 to build Cosinor models). We then used rhythm parameters generated by those models in the correlation analysis. We 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.



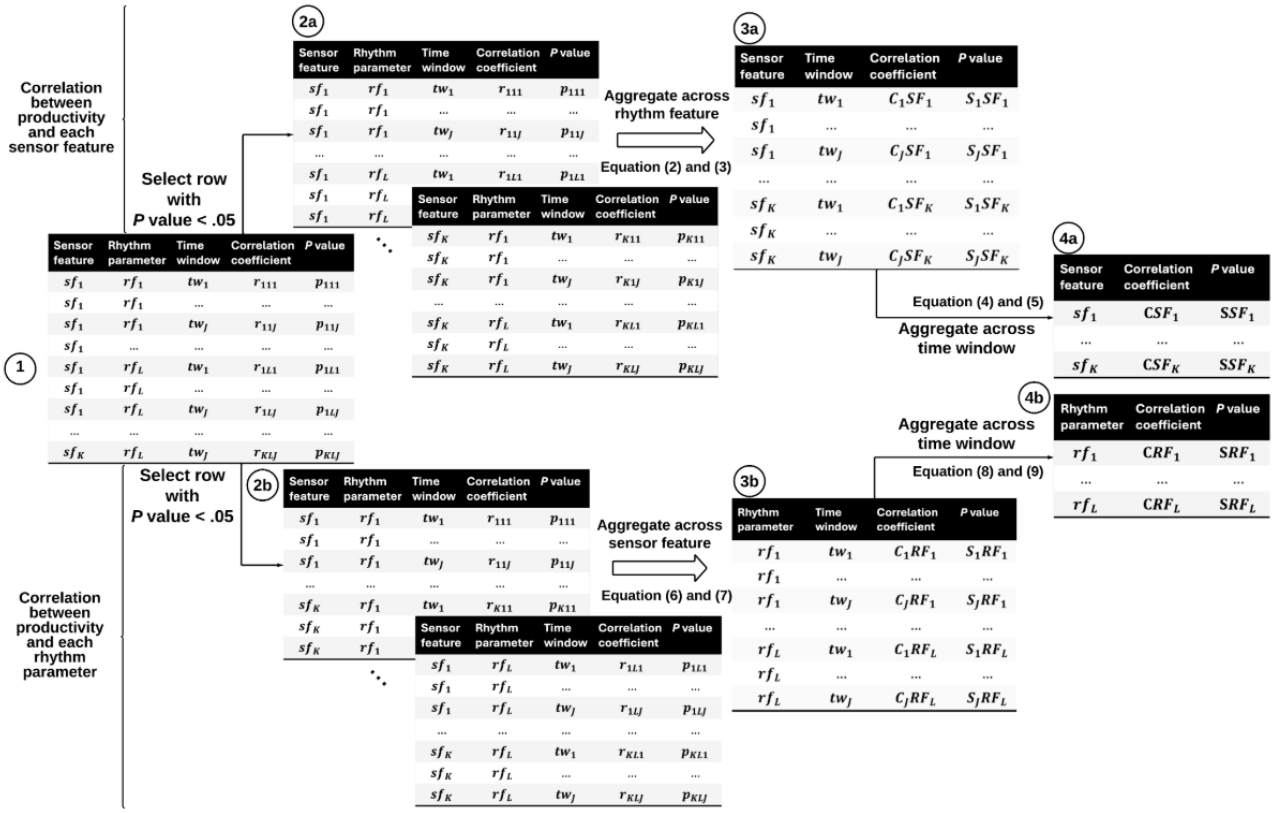
Measuring the Relationship Between Rhythms and Productivity

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.

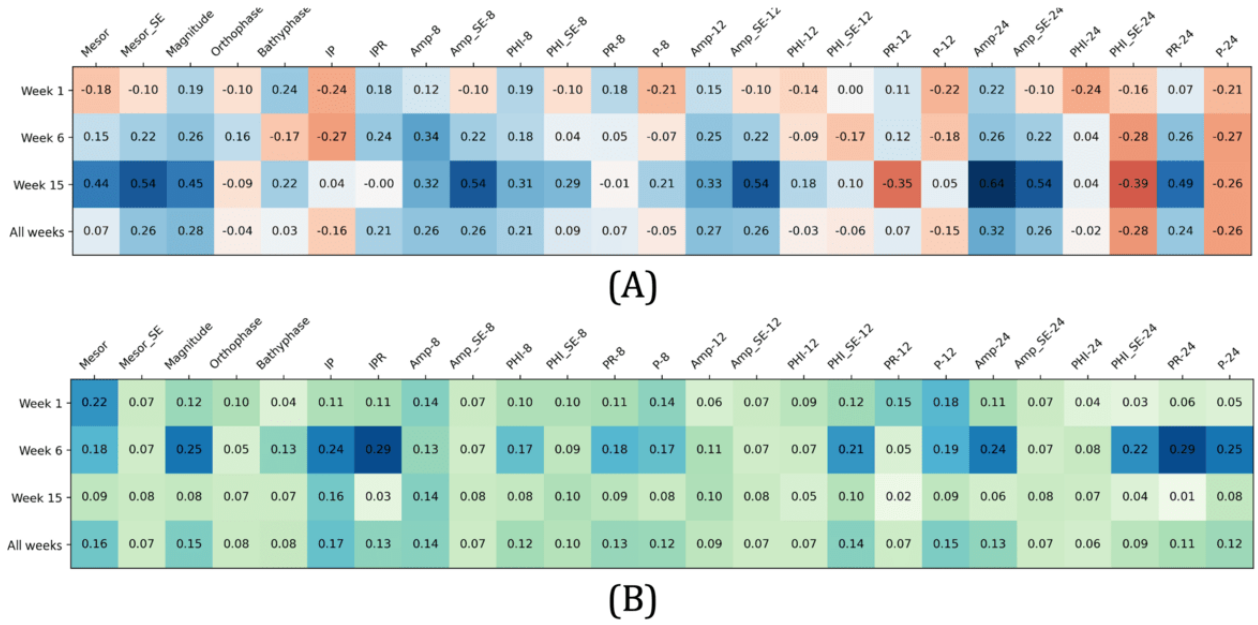
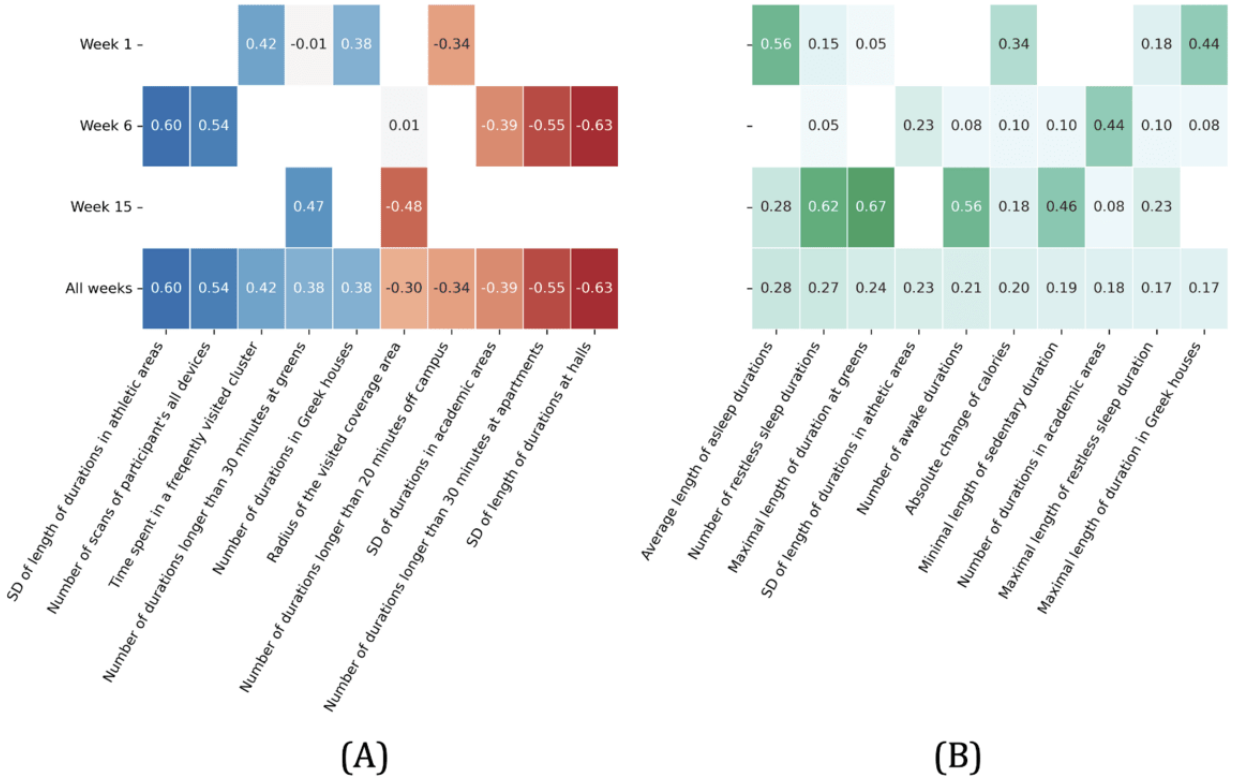


Figure 5. The heat map displays the correlation between sensor features and productivity by week. The left side shows the 10 sensor features with the highest aggregated correlation over all 3 weeks, and the right side shows the 10 sensor features with the highest aggregated significance score over all 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

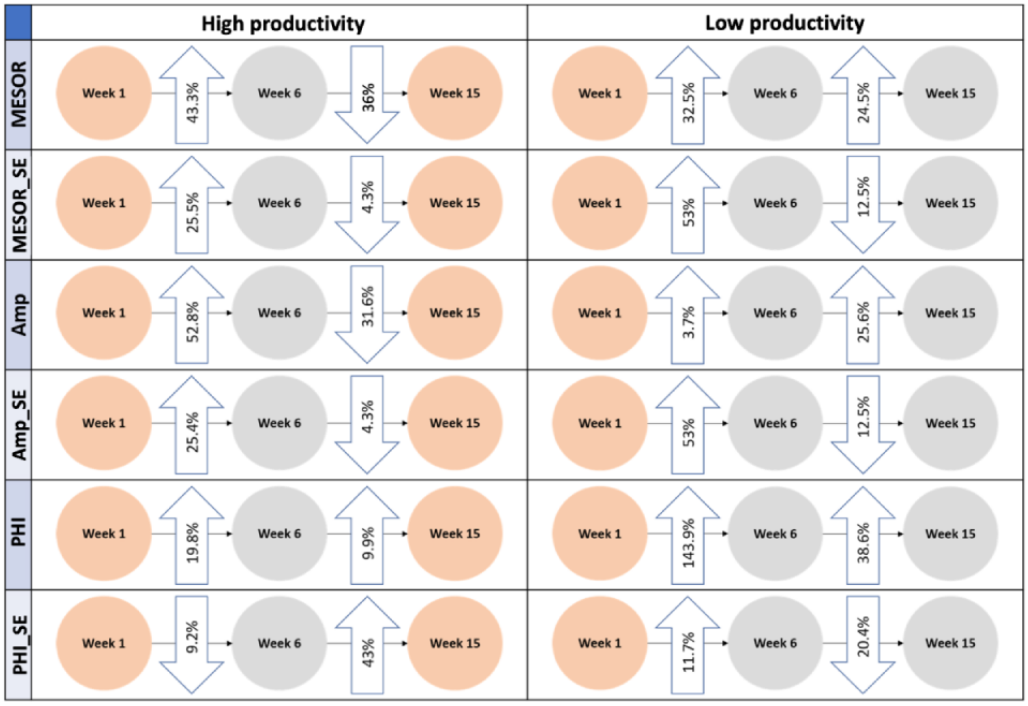
In week 6, the variance of the length or number of stays in academic areas, halls, and apartments showed high negative aggregated correlations with productivity (the left side of Figure 5), indicating that highly productive students had a stable living and studying environment at home and school. Conversely, 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 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 Tesserae [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 Tesserae [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]
19. Murnane 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]
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]
21. Garza JL, Cavallari JM, Eijckelhof 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] [Medline: 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]
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]
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]
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]
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]
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]
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] [Medline: 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]
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]
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]
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]
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] [Medline: 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]
37. Halberg F. Chronobiology. Annu Rev Physiol 1969;31:675-726. [doi: 10.1146/annurev.ph.31.030169.003331] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 18940774]
41. Gierke CL, Cornelissen G. Chronomics Analysis Toolkit (CATkit). Biol Rhythm Res 2016;47(2):163-181. [doi: 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 Tesserae 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]
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] [Medline: 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
PMID:

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 Quttainah[1], PhD; Vinaytosh Mishra[2], PhD; Somayya Madakam[3], PhD; Yotam Lurie[4], PhD; Shlomo Mark[5], PhD

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

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

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

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

[5]Department 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 comprehendible 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.

(JMIR AI 2024;3:e51834) doi:10.2196/51834

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 finalize 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

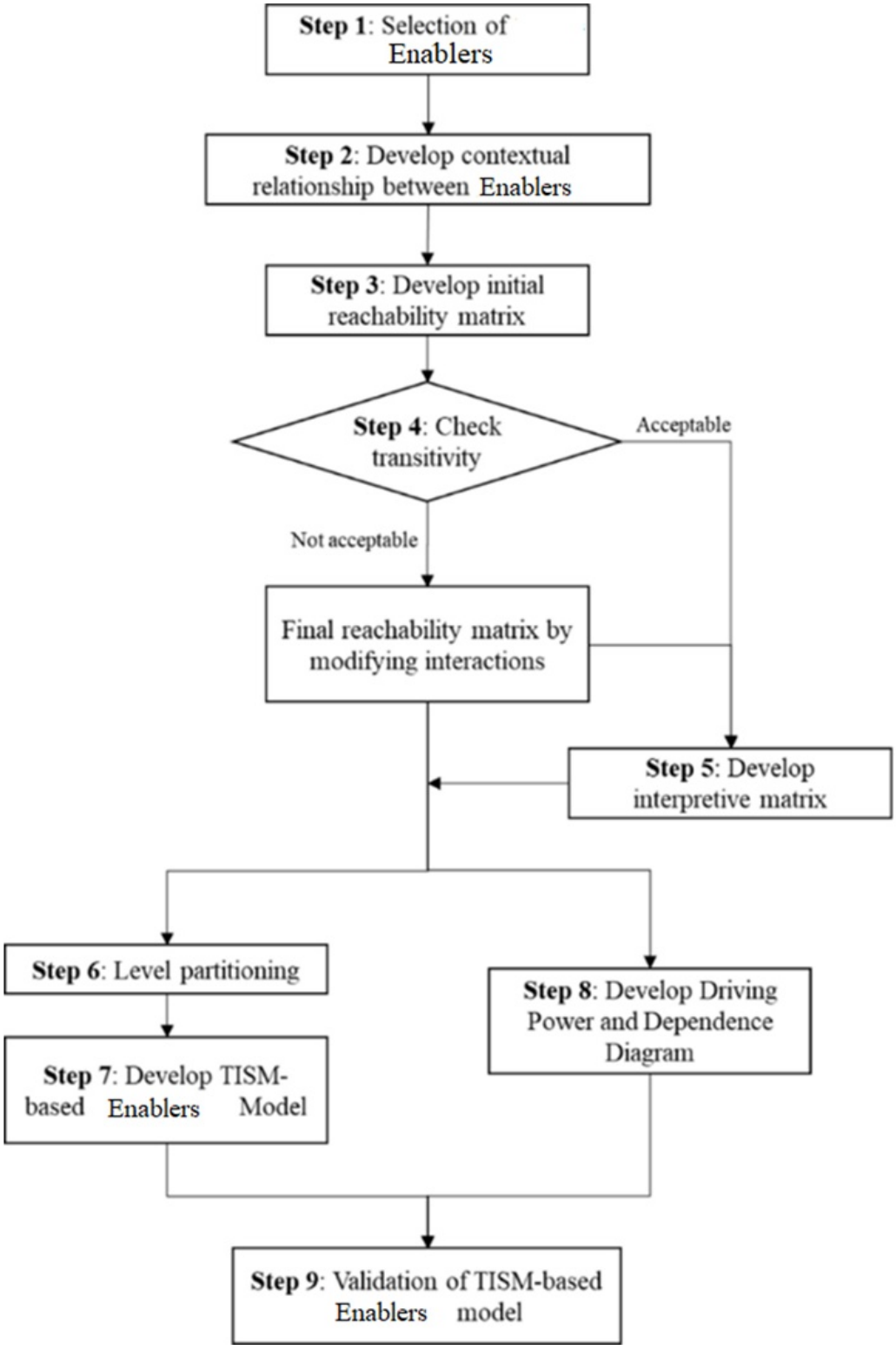
Finally, to address the third research objective, we investigated the relationships among key enablers to inform the development of a suitable medical education LLM. A qualitative research design is useful to understand a phenomenon under study rather than assessing the strength and direction of causal 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

[a]Numbers 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)

λ_{max} = 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)

λ_{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

[a]Y: existence of a relationship between two enablers.

[b]N: no relationship exists between two enablers.

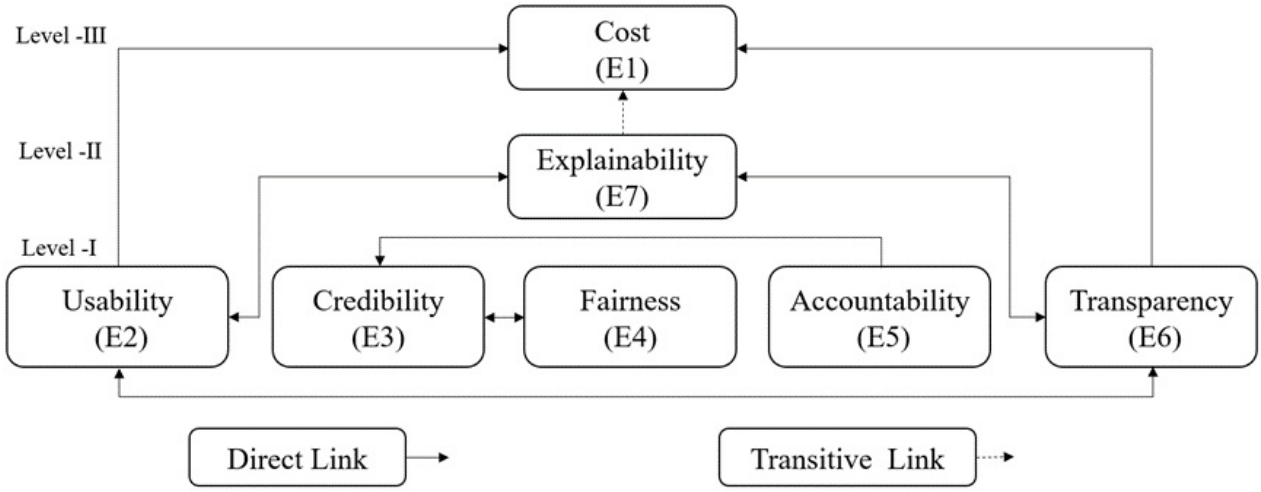
Table 7. Final reachability matrix of the enablers for developing large language models in medical education.

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

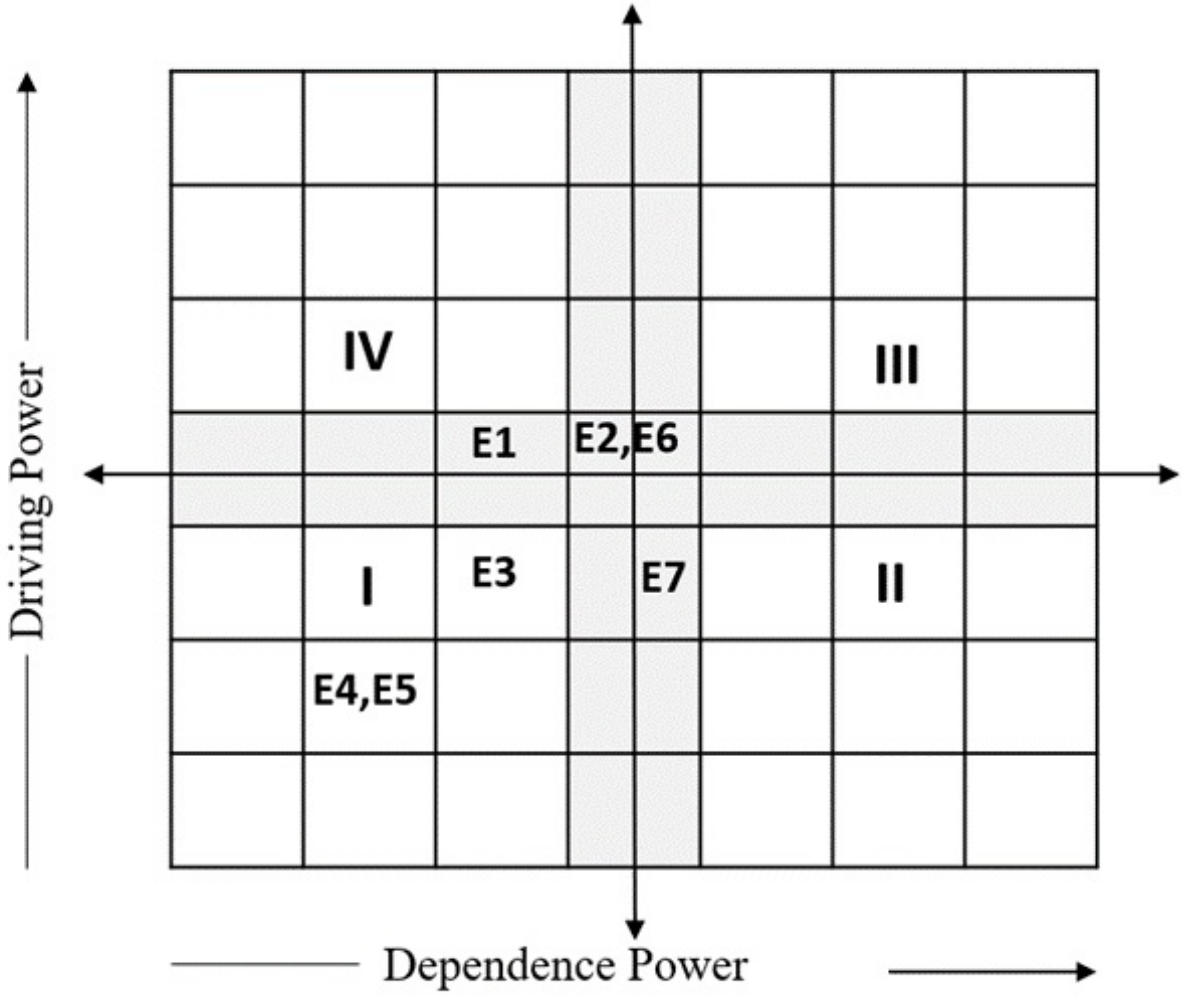
Figure 2. Diagraph of the total interpretive structural model for the development of large language models in medical education.



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.

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.

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].

Authors' Contributions

Conceptualization: VM, MQ, S Madkam, YL, and S Mark; Data curation: VM, S Madakam; 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] [Medline: 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]
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] [Medline: 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]
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] [Medline: 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]
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]
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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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]
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]
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]
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]

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] [Medline: 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]
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] [Medline: 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] [Medline: 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]
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] [Medline: 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]
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]
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]
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]
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]
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]
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]
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]
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]
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] [Medline: 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]
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]
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]
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] [Medline: 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 Artific Intell 2023;5:100152. [doi: 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
PMID:38875562

Original Paper

Learning From International Comparators of National Medical Imaging Initiatives for AI Development: Multiphase Qualitative Study

Kassandra Karpathakis[1], BA, BSc, MPH; Emma Pencheon[2], BA, MBBS, MSc; Dominic Cushnan[3], MBA

[1]Decimal.health, Boston, MA, United States

[2]Foreign, Commonwealth and Development Office, UK Government, London, United Kingdom

[3]NHS 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

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'a 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. Cushnan 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] [Medline: 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. Intercoder reliability in qualitative research: debates and practical guidelines. Int J Qual Methods 2020 Jan 22;19:160940691989922 [FREE Full text] [doi: 10.1177/1609406919899220]
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]

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] [Medline: 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. Raviscioni 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:

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 Assad[4], MD; Sunita Lakhani[5], MD; Maritsa Odabashian[3,6], Bsc (c); Sharon McGee[7,8*], MD, PhD

[1]Department of Oncology, Barbara Ann Karmanos Cancer Institute, Wayne State University, Detroit, MI, United States

[2]Department of Medicine, Division of Internal Medicine, The Ottawa Hospital and the University of Ottawa, Ottawa, ON, Canada

[3]Mary A Rackham Institute, University of Michigan, Ann Arbor, MI, United States

[4]Department of Plastic Surgery, University of Pittsburgh Medical Center, Pittsburgh, PA, United States

[5]Department of Medicine, Division of Internal Medicine, Jefferson Abington Hospital, Philadelphia, PA, United States

[6]The Ottawa Hospital Research Institute, Ottawa, ON, Canada

[7]Department of Medicine, Division of Medical Oncology, The Ottawa Hospital and the University of Ottawa, Ottawa, ON, Canada

[8]Cancer 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

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].

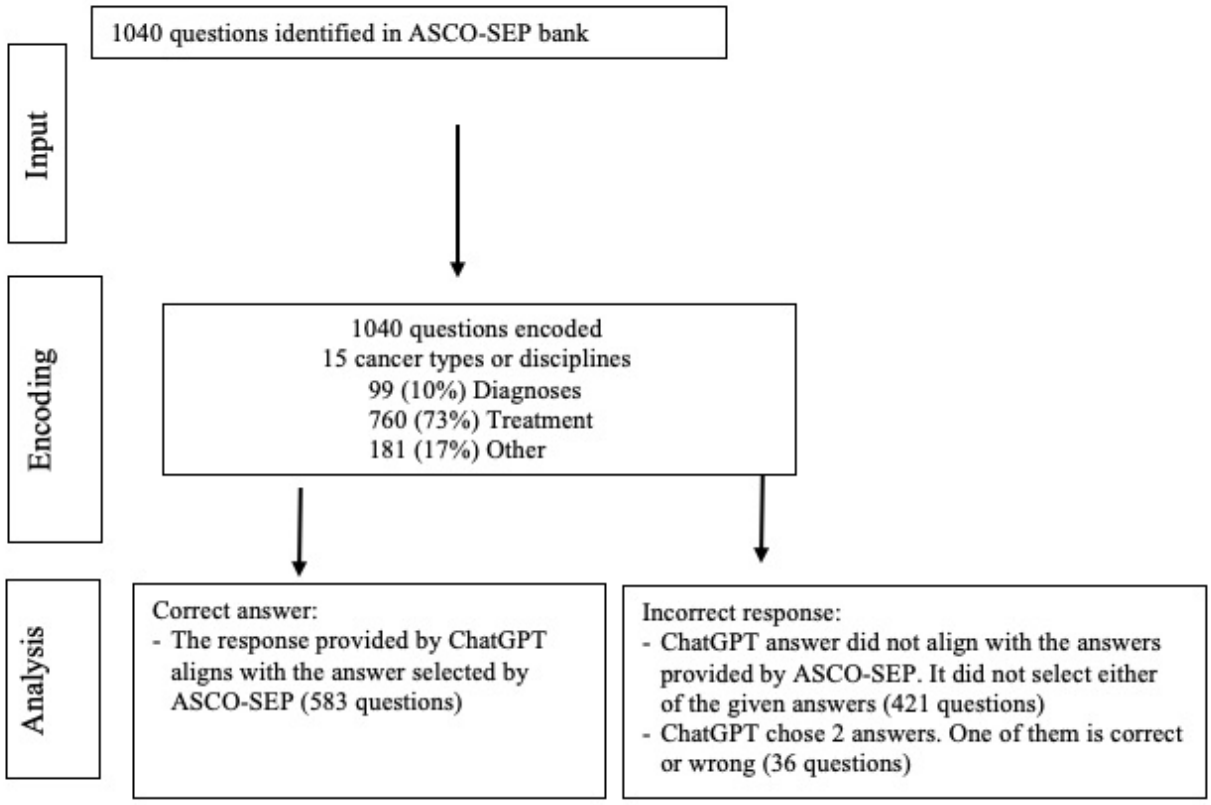
However, the evolving complexities of oncological care necessitate additional tools that can aid oncologists in clinical decision-making. By assessing ChatGPT-3's ability to answer ASCO-SEP questions, this study's objective is to understand ChatGPT's potential to serve as a supportive instrument in clinical 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 ≤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 (%)	P 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

[a]Chi-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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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]
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]
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] [Medline: 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] [Medline: 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] [Medline: 36737653]
17. ASCO-SEP®. American Society of Clinical Oncology® Store. 2023. URL: https://shop.asco.org/asco-sep/#:~:text=ASCO%2DSEP%C2%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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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
PMID:

Original Paper

Role of Ethics in Developing AI-Based Applications in Medicine: Insights From Expert Interviews and Discussion of Implications

Lukas Weidener[1], BSc, Dr med; Michael Fischer[1], 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

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

[a]AI: 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.

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.

Textbox 1. Overview of the 3 main categories with a total of 7 subcategories from the analysis of interviews with experts in artificial intelligence.

Essential foundation

- Awareness
- Consequences
- Data protection

Results in the foreground

- Performance
- Economic efficiency

Obstacle to progress

- 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.

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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 36853242]
11. Scerri M, Grech V. Artificial intelligence in medicine. Early Hum Dev 2020 Jun;145:105017. [doi: 10.1016/j.earlhumdev.2020.105017] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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]
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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 34550305]
28. Hedlund M, Persson E. Expert responsibility in AI development. AI Soc 2022 Jun 13. [doi: 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]
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]
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]
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]
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]
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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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/f7cfbf5a-19fa-4a9e-a3cc-114ec8b529e4 [doi: 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] [Medline: 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]
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] [Medline: 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]
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] [Medline: 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] [Medline: 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] [Medline: 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]

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
PMID:38875585

Original Paper

Identifying Symptoms Prior to Pancreatic Ductal Adenocarcinoma Diagnosis in Real-World Care Settings: Natural Language Processing Approach

Fagen Xie[1], PhD; Jenny Chang[1], MPH; Tiffany Luong[1], MPH; Bechien Wu[1], MD, MPH; Eva Lustigova[1], MPH; Eva Shrader[2], MS; Wansu Chen[1], PhD

[1]Department of Research and Evaluation, Kaiser Permanente Southern California, Pasadena, CA, United States

[2]Pancreatic 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

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.

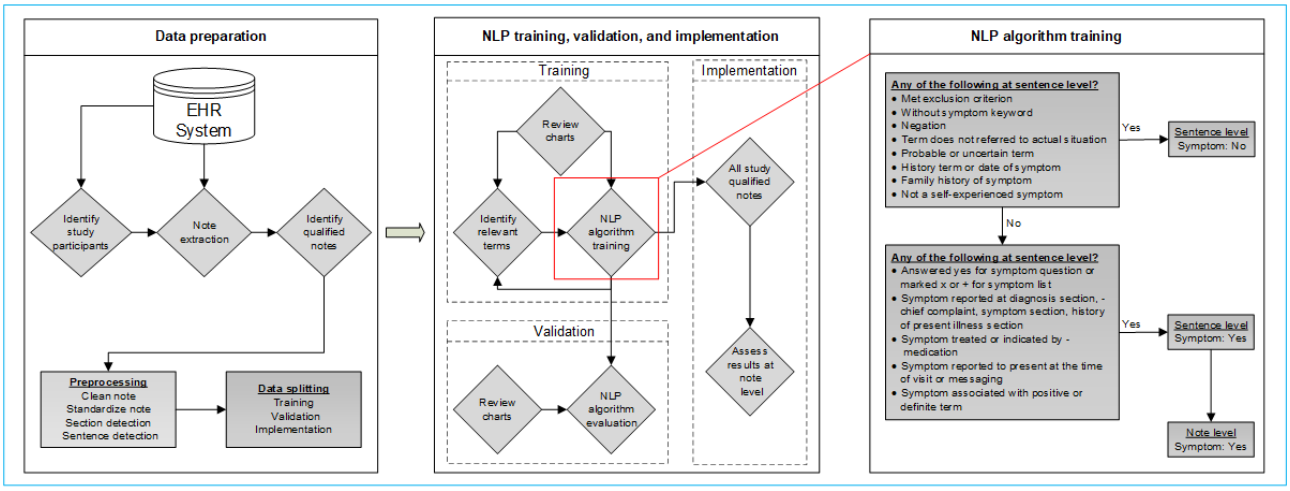
Ethical Considerations

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 × PPV × 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)

[a]PDAC: 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

[a]TP: true positive.

[b]TN: true negative.

[c]FP: false positive.

[d]FN: false negative.

[e]PPV: positive predicted value.

[f]NPV: Negative predicted value.

[g]Hospital encounter notes were excluded with the exception of emergency notes.

[h]DVT: 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)

[a]PDAC: pancreatic ductal adenocarcinoma.

[b]DVT: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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]
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] [Medline: 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] [Medline: 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] [Medline: 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]
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]
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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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]
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] [Medline: 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]
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]
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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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]

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] [Medline: 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]
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] [Medline: 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] [Medline: 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]
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. 2022 Presented at: TMTNLP 2022; July 7, 2002; Philadelphia, PA. [doi: 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]
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] [Medline: 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] [Medline: 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] [Medline: 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 Senst, 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
PMID:38875566

Original Paper

Beyond the Hype—The Actual Role and Risks of AI in Today's Medical Practice: Comparative-Approach Study

Steffan Hansen[1], MA; Carl Joakim Brandt[1], PhD; Jens Søndergaard[1], 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. Winsløws 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

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	• 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	• A brief overview of AI in general practice • Lack of in-depth or specific study citations • Offering a more concise perspective

[a]AI: 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	• A total of 23 out of 32 (72%) precise peer-reviewed articles, with high accuracy • A total of 9 nonexistent articles, with specific inaccuracies	• 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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]
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://d4mucfpksywv.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]
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]
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] [Medline: 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]
15. Brandtzaeg P, Følstad A. Why People Use Chatbots. Cham, Switzerland: Springer Link; 2017:377-392.

Abbreviations

AI: artificial intelligence
LAMBADA: LAnguage 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
PMID:

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 Zhang[2], BE; Yi Xiao[2], PhD; Yingyu Wang[2], BE

[1]State Key Laboratory of Communication Content Cognition, People's Daily Online, Beijing, China

[2]School 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

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?

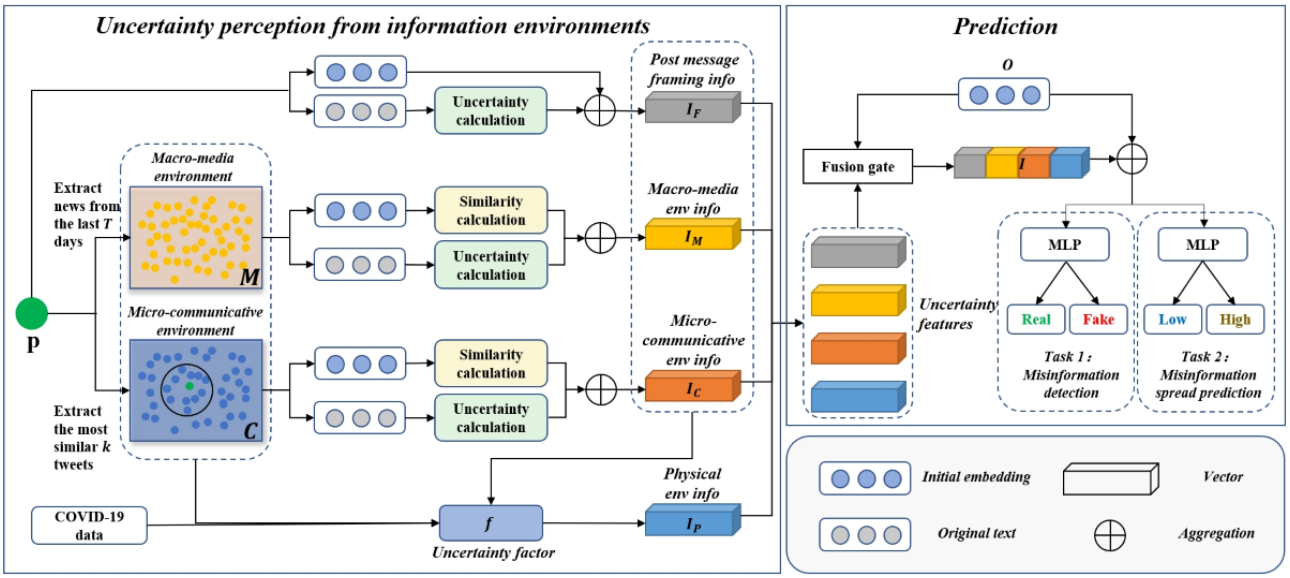
Methods

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_u, 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:

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:

$$F^{A,\text{Mean}}(s)=\text{Mean}(\text{Norm}(\{F_i(s)\}^{|F|}_{i=1})) \quad (4)$$

where Norm(·) 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:

$$M = \{\boldsymbol{e}: \boldsymbol{e} \in E, 0 \le t_p - t_e \le T\} \quad (5)$$

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 $\boldsymbol{p}$ and $\boldsymbol{e}$, 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 $\boldsymbol{p}$ and each news item $\boldsymbol{e}$ 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 $\boldsymbol{p}$ to the environment:

where $\boldsymbol{e}^{M}_{i}$ represents each news item in M and  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:

$$I_M = \text{MLP}(S_M \oplus U_M) \quad (9)$$

where  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:

$$C' = \{v : v \in V, 0 \le t_p - t_v \le T\} \quad (10)$$

where V denotes the set of all collected tweet items and t_v represents the release time of the tweet v.

$$C = \{v : v \in \text{Top}k(p, C')\} \quad (11)$$

where Top$k(\cdot)$ 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:

$$I_F = \text{MLP}[\boldsymbol{F}(p) \oplus \boldsymbol{p}] \quad (15)$$

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:

$$f^{ph}_{i} = \text{Norm}(\log(1+\text{abs}(N_i^{\text{Cases}} - N_{i-1}^{\text{Cases}})) \times \log(1+\text{abs}(N_i^{\text{News}} - N_{i-1}^{\text{News}}))) \quad (16)$$

where f^{ph}_{i} 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 $f^{ph}(p)$.

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:

$$I_P = (f^{ph} \times \text{I}_\text{M}) \oplus (f^{ph} \times \text{I}_\text{C}) \oplus (f^{ph} \times \text{I}_\text{F}) \quad (17)$$

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:

$$I_{\text{EUP}} = I_M \oplus I_C \oplus I_F \oplus I_P \quad (18)$$

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:

$$I=g_M I_M + g_C I_C + g_F I_F + g_P I_P \quad (20)$$

where $\boldsymbol{o}$ denotes the last-layer feature from the misinformation baseline algorithm. The gating vector g_M=sigmoid(Linear($\boldsymbol{o}I_M$)) and g_C, g_F, and g_P are obtained in the same way. Then, we concatenated $\boldsymbol{o}$ 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

[a]News items in M=58,095. The corresponding mean and range are 988 and 10-2511, respectively.

[b]Tweet 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.

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.

Textbox 1. Baseline models.

1. **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.

2. **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).

3. **BERT**

   Bidirectional encoder representations from transformers (BERT) [52] is a pretrained language model based on deep bidirectional transformers.

4. **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.

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	$F_{1\ \text{fake}}$	$F_{1\ \text{real}}$
EUP	*0.753*	*0.739*	*0.800*	*0.677*
BiLSTM[c]	0.733	0.729	0.783	0.683
BiLSTM + EUP	*0.755*	*0.743*	*0.798*	*0.688*
$EANN_T$[d]	0.745	0.730	0.795	0.664
$EANN_T$ + EUP	*0.767*	*0.765*	*0.806*	*0.708*
BERT[e]	0.755	0.743	*0.797*	0.689
BERT + EUP	*0.771*	*0.767*	0.796	*0.738*
BERT-Emo	0.749	0.740	0.789	0.691
BERT-Emo + EUP	*0.768*	*0.763*	*0.799*	*0.726*

[a]EUP: Environmental Uncertainty Perception.

[b]The best result in each group is in italics.

[c]BiLSTM: bidirectional long short-term memory.

[d]$EANN_T$: event adversarial neural networks.

[e]BERT: 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	$F_{1\ \text{low}}$	$F_{1\ \text{high}}$
EUP	*0.710*	*0.710*	*0.719*	*0.701*
BiLSTM[c]	0.707	0.705	0.684	0.726
BiLSTM + EUP	*0.734*	*0.733*	*0.738*	*0.729*
$EANN_T$[d]	0.717	0.716	0.734	0.698
$EANN_T$ + EUP	*0.726*	*0.726*	*0.736*	*0.716*
BERT[e]	0.728	0.728	0.728	0.728
BERT + EUP	*0.743*	*0.743*	*0.752*	*0.734*
BERT-Emo	0.733	0.733	0.730	0.737
BERT-Emo + EUP	*0.741*	*0.741*	*0.733*	*0.749*

[a]EUP: Environmental Uncertainty Perception.

[b]The best result in each group is in italics.

[c]BiLSTM: bidirectional long short-term memory.

[d]$EANN_T$: event adversarial neural networks.

[e]BERT: 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	$F_{1\ \text{fake}}$	$F_{1\ \text{real}}$
EUP[b]	*0.753*	*0.739*	0.800	0.677
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	*0.755*	*0.743*	*0.798*	*0.688*
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	*0.771*	*0.767*	*0.796*	*0.738*
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

[a]The best result in each group is in italics.

[b]EUP: Environmental Uncertainty Perception.

[c]BiLSTM: bidirectional long short-term memory.

[d]BERT: 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]	*0.710*	*0.710*	0.719	*0.701*
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	*0.721*	0.692
BiLSTM[c] **+ EUP**	*0.734*	*0.733*	0.738	*0.729*
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	*0.753*	0.693
BERT[d] **+ EUP**	*0.743*	*0.743*	0.752	*0.734*
Without I_M	0.741	0.739	0.764	0.713
Without I_C	0.741	0.738	*0.766*	0.711
Without I_F	0.736	0.735	0.753	0.716
Without I_P	0.740	0.738	0.759	0.717

[a]The best result in each group is in italics.

[b]EUP: Environmental Uncertainty Perception.

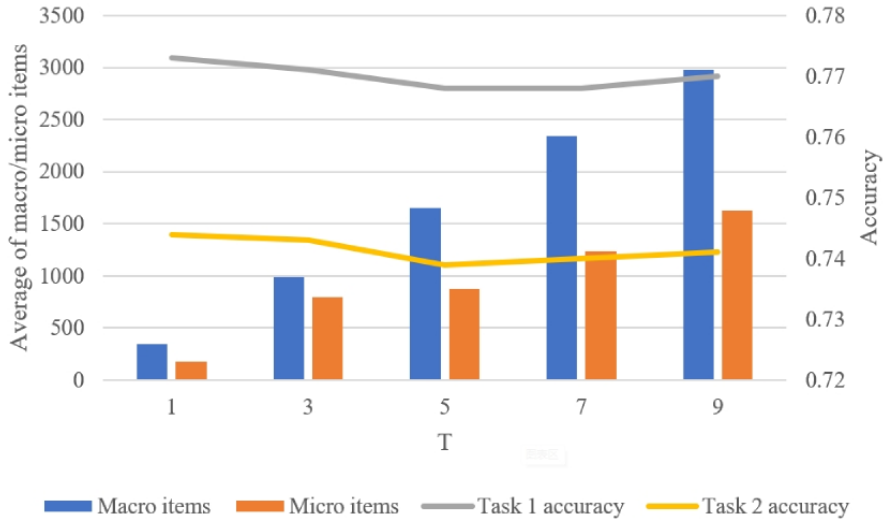
[c]BiLSTM: bidirectional long short-term memory.

[d]BERT: 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.

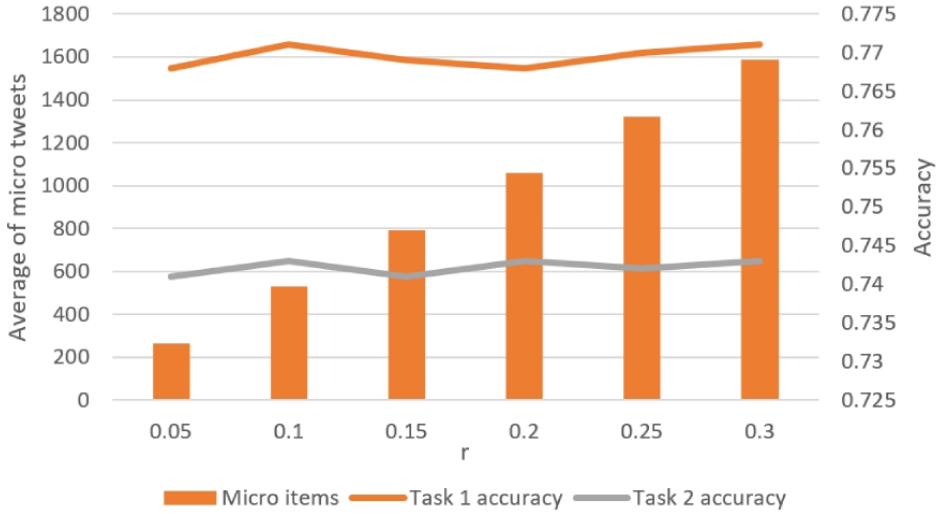
Figure 2. The effect of the day parameter T. Lines show the accuracies of both tasks and bars show the average number of news and tweet items in the environments.



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.

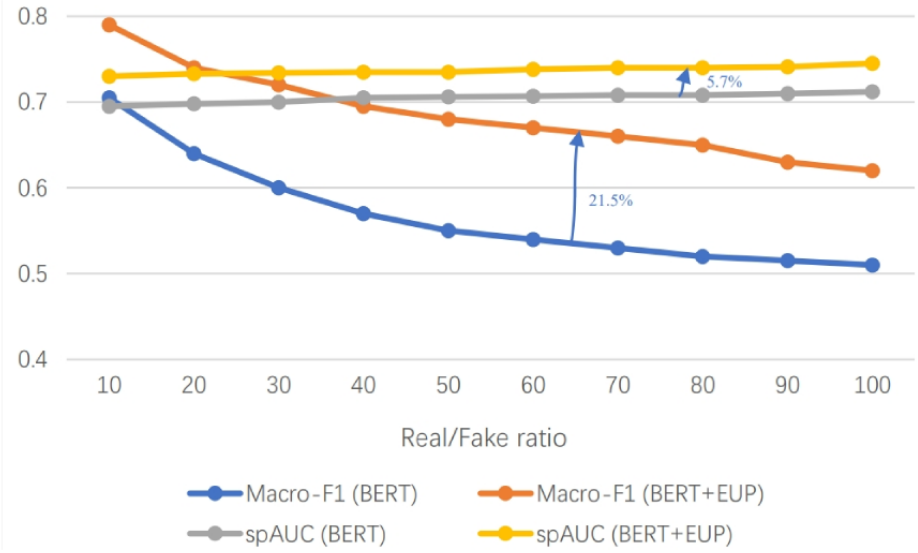
Figure 3. The effect of the rate parameter r. Lines show the accuracies of both tasks and bars show the average number of tweet items in the environment.



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≤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≤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] [Medline: 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]
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] [Medline: 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]
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]
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]
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]
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] [Medline: 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]
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]
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]
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]
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]

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]
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]
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] [Medline: 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] [Medline: 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] [Medline: 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]
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]
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]
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]
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]
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]
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]
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]
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]
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]
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]
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] [Medline: 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]

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]
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] [Medline: 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] [Medline: 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]
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] [Medline: 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] [Medline: 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] [Medline: 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]
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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 33103600]
47. Higgins-Dunn N. A dog in Hong Kong tests positive for the coronavirus, WHO officials confirm. CNBC. 2022. URL: https://www.cnbc.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] [Medline: 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]
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] [Medline: 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]
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]
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]
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]
57. Hartigan JA, Wong MA. Algorithm AS 136: a K-means clustering algorithm. Applied Statistics 1979;28(1):100. [doi: 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] [Medline: 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] [Medline: 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]
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]
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] [Medline: 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]
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]
66. McClish DK. Analyzing a portion of the ROC curve. Med Decis Making 1989;9(3):190-195. [doi: 10.1177/0272989X8900900307] [Medline: 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]
68. Brashers D. Communication and uncertainty management. Journal of Communication 2001;51(3):477-497. [doi: 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 Yiannakoulias; 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
PMID:38875583

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 Pan[1], MPhil; Hao Luo[2], PhD; Gary Cheung[3], PhD; Huiquan Zhou[2], PhD; Reynold Cheng[1], PhD; Sarah Cullum[3], PhD; Chuan Wu[1], PhD

[1]Department of Computer Science, The University of Hong Kong, Hong Kong, China (Hong Kong)

[2]Department of Social Work and Social Administration, The University of Hong Kong, Hong Kong, China (Hong Kong)

[3]Department 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 ≤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

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

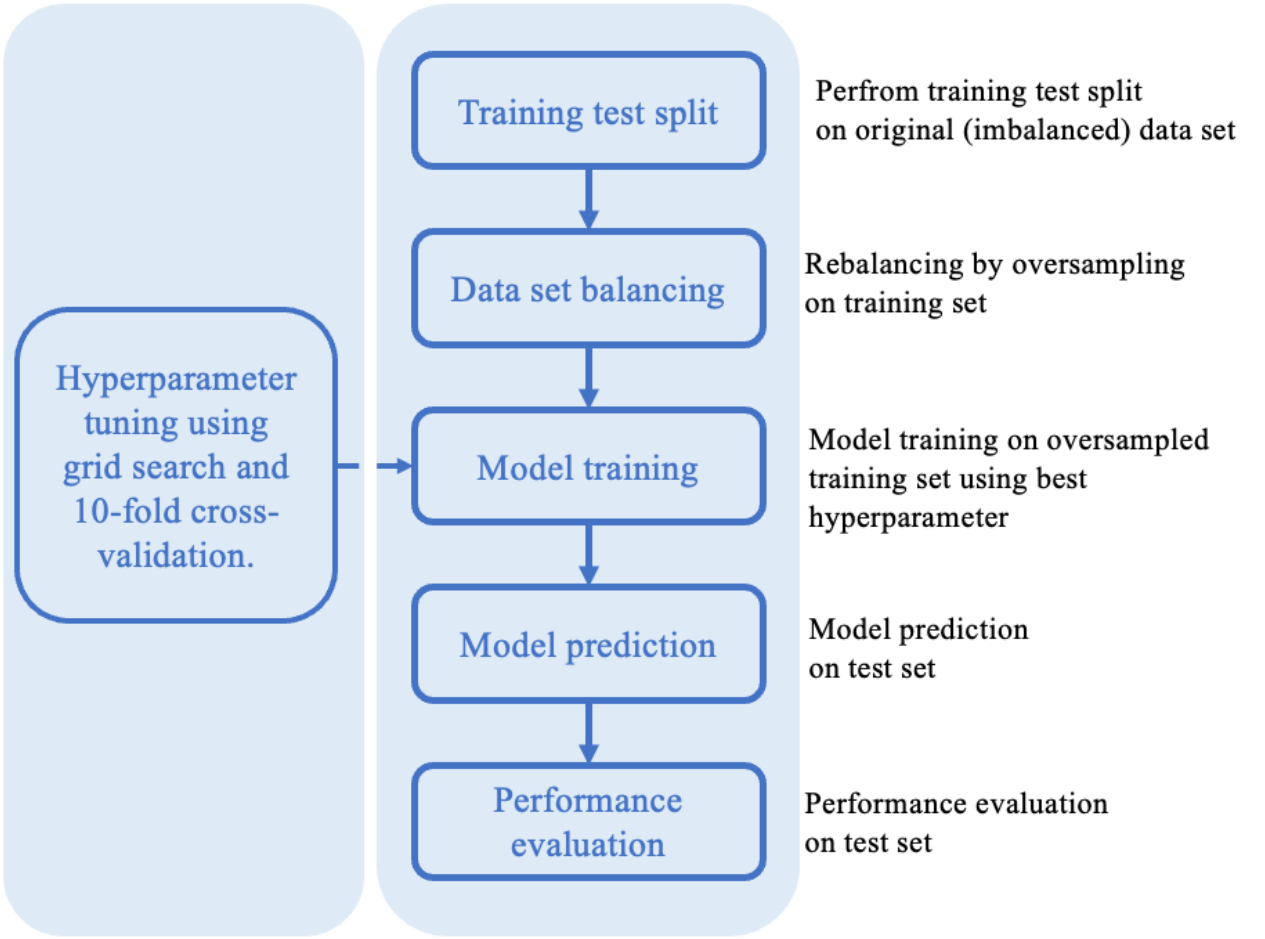
We applied 3 state-of-the-art machine learning models and regularized logistic regression to predict 6-month and 12-month mortality using the features available from interRAI-HC. The RF is a machine learning algorithm that uses decision trees [37]. The RF provides highly accurate predictions with a very 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)

[a]HC: 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.

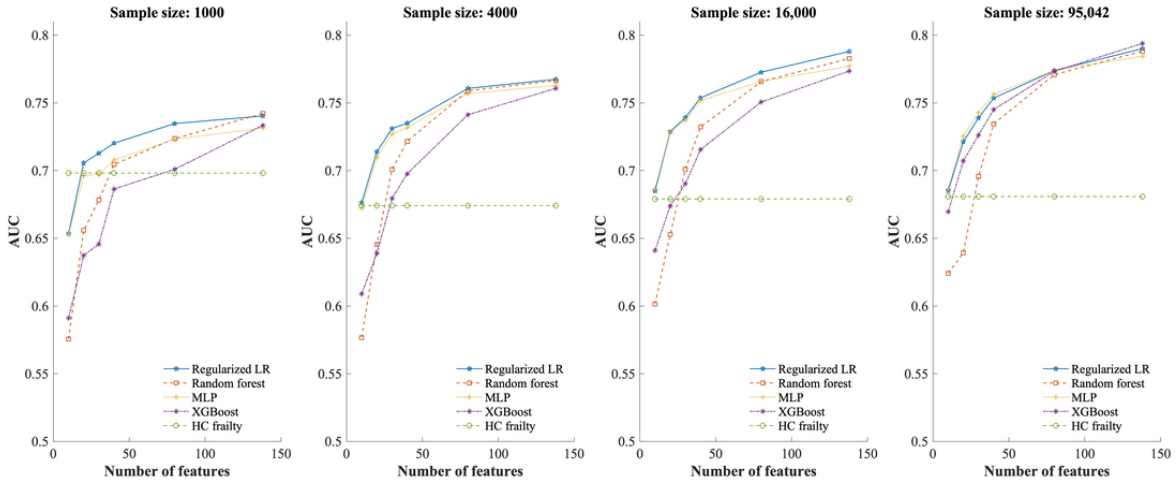


Figure 3. Average sensitivities of classifiers and frailty scale for 6-month mortality prediction on balanced data set. HC: home care; LR: logistic regression; MLP: multilayer perceptron; XGBoost: extreme gradient boosting.

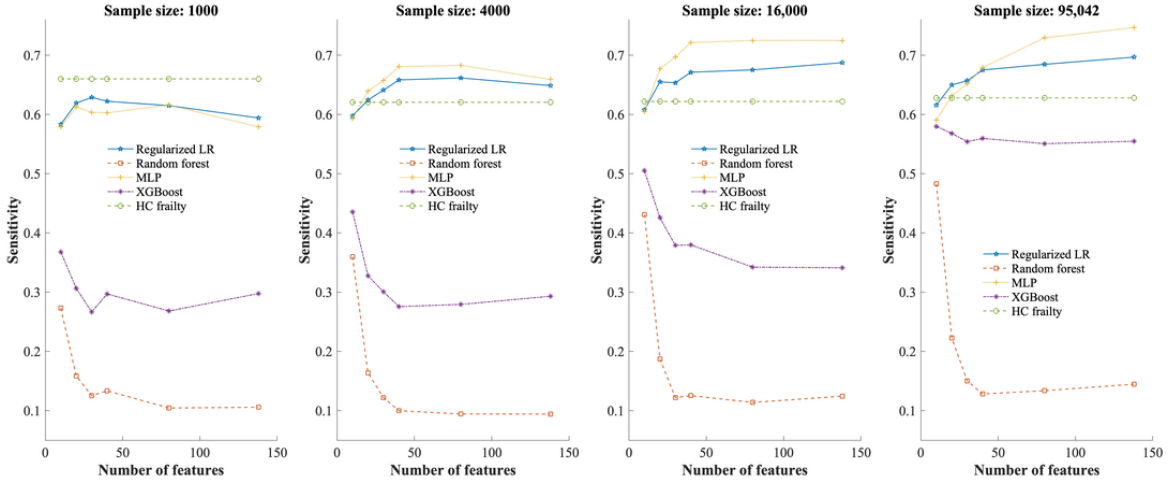


Figure 4. Average specificities of classifiers and frailty scale for 6-month mortality prediction on balanced data set. HC: home care; LR: logistic regression; MLP: multilayer perceptron; XGBoost: extreme gradient boosting.

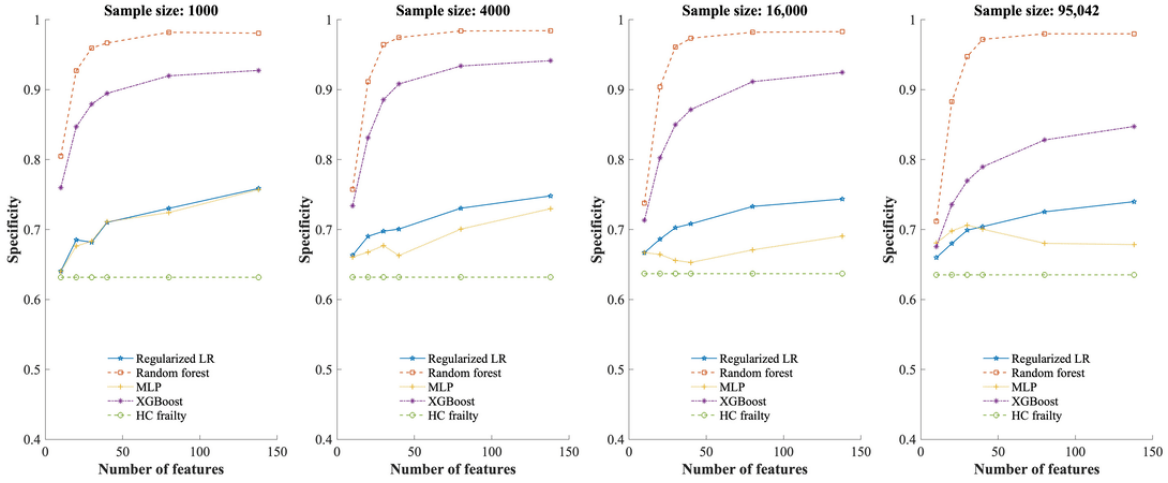
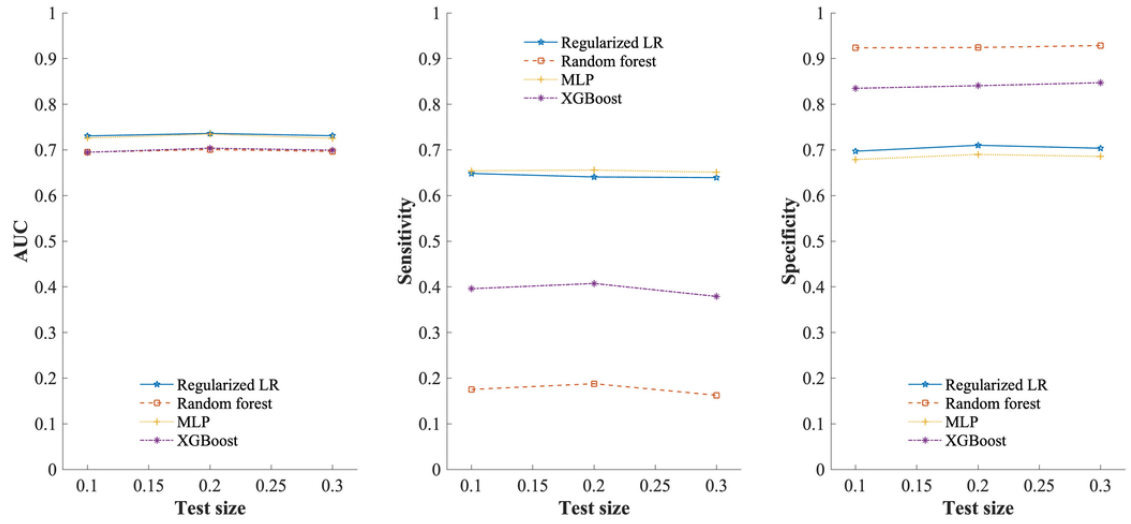


Figure 5. Average AUCs, sensitivities, and specificities of frailty scales for 6-month mortality prediction by test sizes on balanced data set. AUC: area under the curve; LR: logistic regression; MLP: multilayer perceptron; XGBoost: extreme gradient boosting.



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] [Medline: 24666801]
2. Rodríguez-Mañas L, Féart C, Mann G, Viña J, Chatterji S, Chodzko-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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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/fpsyt.2019.00926] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 31075167]
33. Espaulella J, Arnau A, Cubí D, Amblàs J, Yánez 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 33437945]
37. Sternberg SA, Schwartz AW, Karunananthan 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] [Medline: 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]
40. Colombet I, Ruelland A, Chatellier G, Gueyffier F, Degoulet P, Jaulent 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]
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]
42. Hoerl AE, Kennard RW. Ridge regression: biased estimation for nonorthogonal problems. Technometrics 1970;12(1):55-67. [doi: 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]
44. Viloria 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]
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] [Medline: 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] [Medline: 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] [Medline: 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.jjcc.2021.11.006] [Medline: 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] [Medline: 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]

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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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]

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
PMID:

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 Hamwi[2], BA; Oana Bucsea[2], MA; Carol Cheng[3], MSc; Hang Du[4], MA; Lorenzo Fabrizi[5], PhD; Sara Jasim[2], MA; Lesley Johannsson[6], BScN; Laura Jones[5], PhD; Maria Pureza Laudiano-Dray[5], MRes; Judith Meek[7], MBBS, PhD; Neelum Mistry[5], MSc; Vibhuti Shah[8], MSc, MD; Ian Stedman[9], LLB, PhD; Xiaogang Wang[4], PhD; Rebecca Pillai Riddell[2], PhD

[1]School of Psychology, University of Ottawa, Children's Hospital of Eastern Ontario Research Institute, Ottawa, ON, Canada
[2]Department of Psychology, York University, Toronto, ON, Canada
[3]Department of Nursing, Mount Sinai Hospital, Toronto, ON, Canada
[4]Department of Mathematics and Statistics, York University, Toronto, ON, Canada
[5]Department of Neuroscience, Physiology and Pharmacology, University College London, London, United Kingdom
[6]Mount Sinai Hospital, Toronto, ON, Canada
[7]Neonatal Care Unit, University College London Hospitals, London, United Kingdom
[8]Department of Pediatrics, Mount Sinai Hospital, Toronto, ON, Canada
[9]School 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

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)	—	—

[a]Not 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.

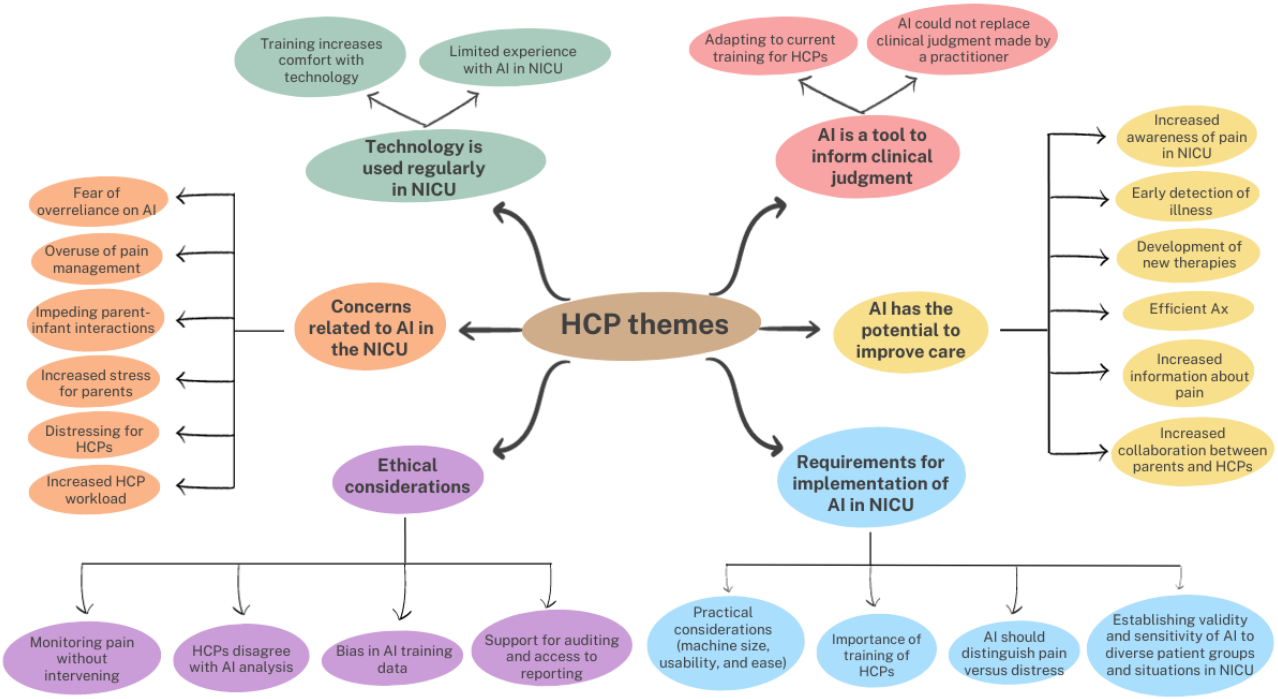


Table 2. Key themes identified by HCPs[a] with regard to the use and integration of AI[b] for pain assessment in the NICU[c].

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.	• "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.	• 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.	• "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.	• 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.	• "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.	• "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."

[a]HCP: health care professional.

[b]AI: artificial intelligence.

[c]NICU: 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.

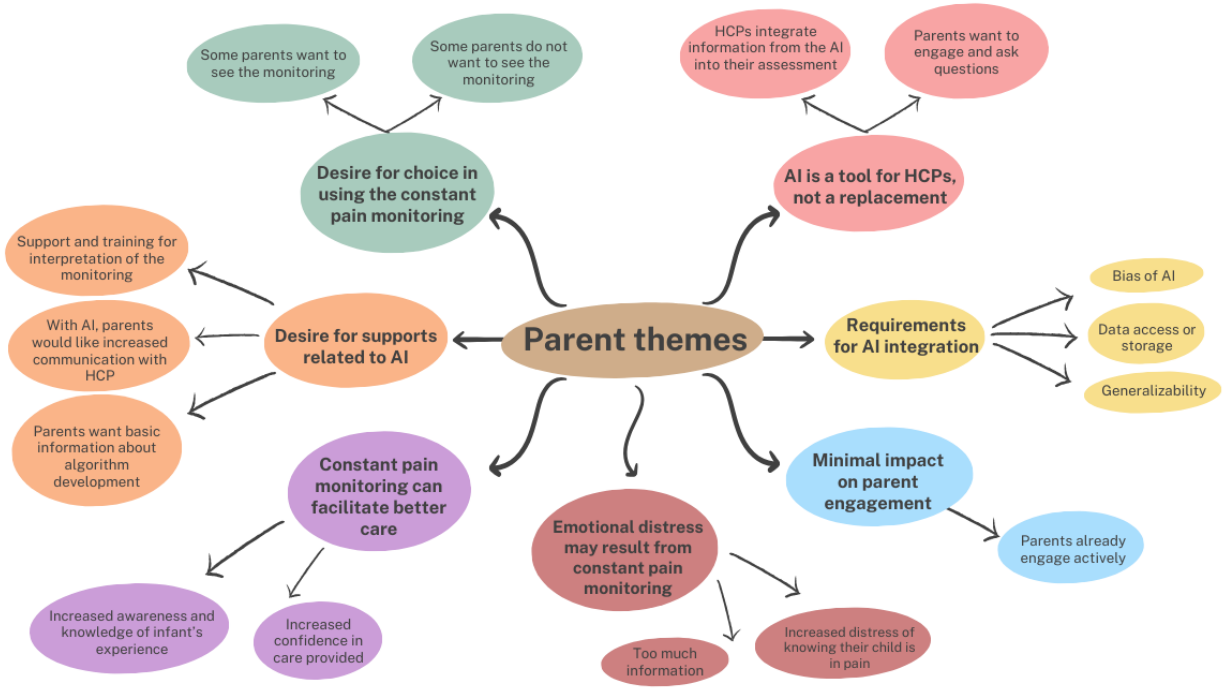


Table 3. Key themes identified by parents with regard to the use and integration of AI[a] for pain assessment in the NICU[b].

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).	• "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.	• "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.	• "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.	• "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.	• "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 HCPs[c], not a replacement	Parents indicated that constant pain monitoring should be used as a tool to inform clinical judgment.	• "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.	• "... 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."

[a]AI: artificial intelligence.

[b]NICU: neonatal intensive care unit.

[c]HCP: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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]
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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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]
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]
27. Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol 2006;3(2):77-101. [doi: 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]
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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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]
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] [Medline: 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 Görges, 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
PMID:38875686

Original Paper

Reidentification of Participants in Shared Clinical Data Sets: Experimental Study

Daniela Wiepert[1], BA; Bradley A Malin[2,3,4], PhD; Joseph R Duffy[1], PhD; Rene L Utianski[1], PhD; John L Stricker[1], PhD; David T Jones[1], MD; Hugo Botha[1], MBChB

[1]Department of Neurology, Mayo Clinic, Rochester, MN, United States

[2]Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN, United States

[3]Department of Biostatistics, Vanderbilt University Medical Center, Nashville, TN, United States

[4]Department 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.

(JMIR AI 2024;3:e52054) doi:10.2196/52054

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

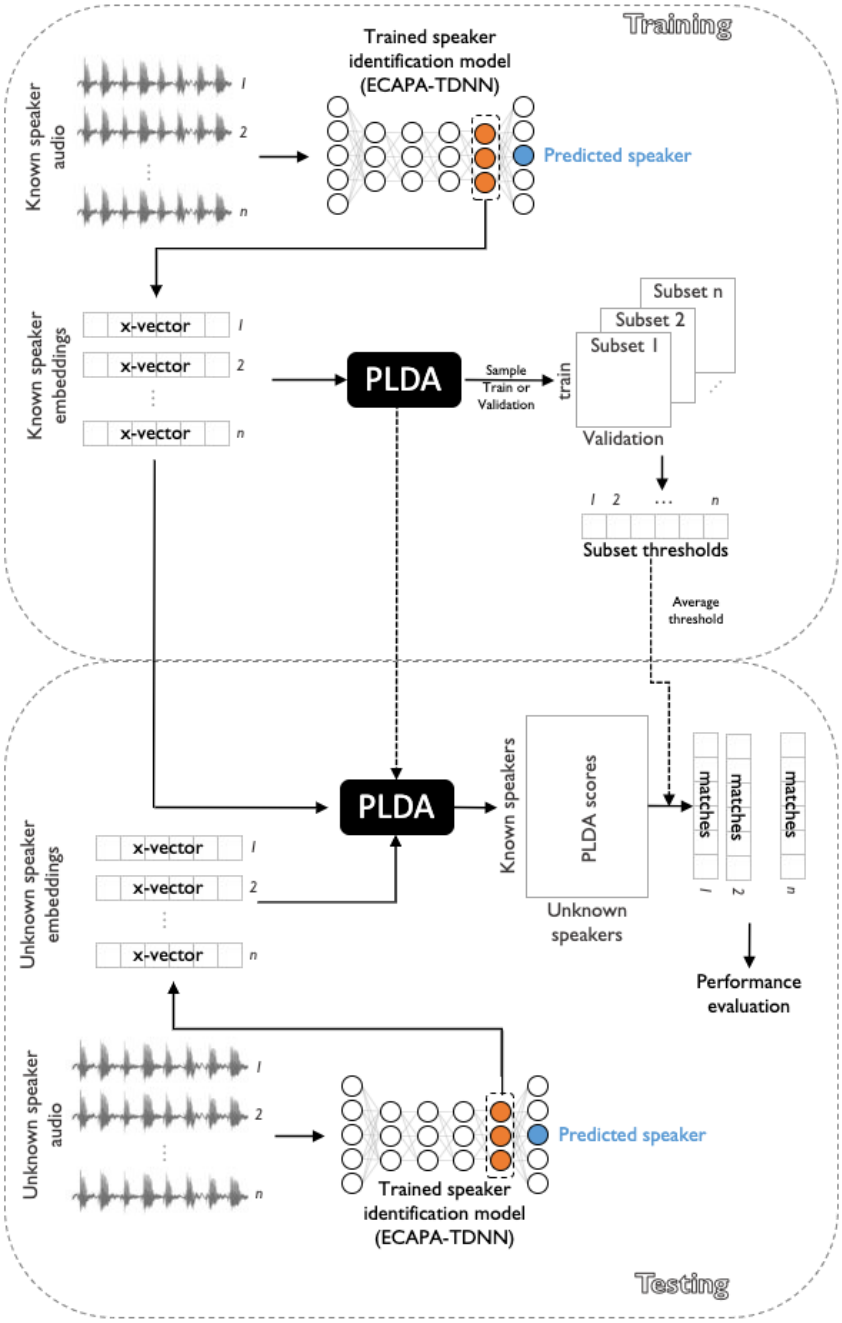
In this study, we addressed the risk of reidentification in a series of experiments exploring the reidentifiability of medical speech recordings without using any metadata. We accomplished this goal by modeling an adversarial attack using a state-of-the-art speaker identification architecture wherein 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)

[a]AMR: alternating motion rate.

[b]SMR: sequential motion rate.

[c]354 total unique speakers.

[d]Samples instead of recordings.

[e]551 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:

$$\text{minDCF} = C_{FR} \times FR \times \text{prior}_{\text{target}} + C_{FA} \times FA \times (1 - \text{prior}_{\text{target}}) \quad \textbf{(1)}$$

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	• 1000 to 7205 (varied known) • 6000 (varied unknown)	• 163 (varied known) • 150 to 1000 (varied unknown)	• 158 (varied known) • 145 to 995 (varied unknown)	• 5	10	0.1	0.001
VoxCeleb: full-overlap worst-case scenario	20	• 1000 to 7205 (varied known) • 6000 (varied unknown)	• 163 (varied known) • 150 to 1000 (varied unknown)	• 0	• 163 (varied known) • 150 to 1000 (varied unknown)	10	0.1	0.001
Mayo Clinic speech recordings: cross-task	20	• 500	• 55	• 50	• 5	1	1	0.01
Mayo Clinic speech recordings: within task	20	• 500[a]	• 55	• 50	• 5	1	1	0.01

[a]Word 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

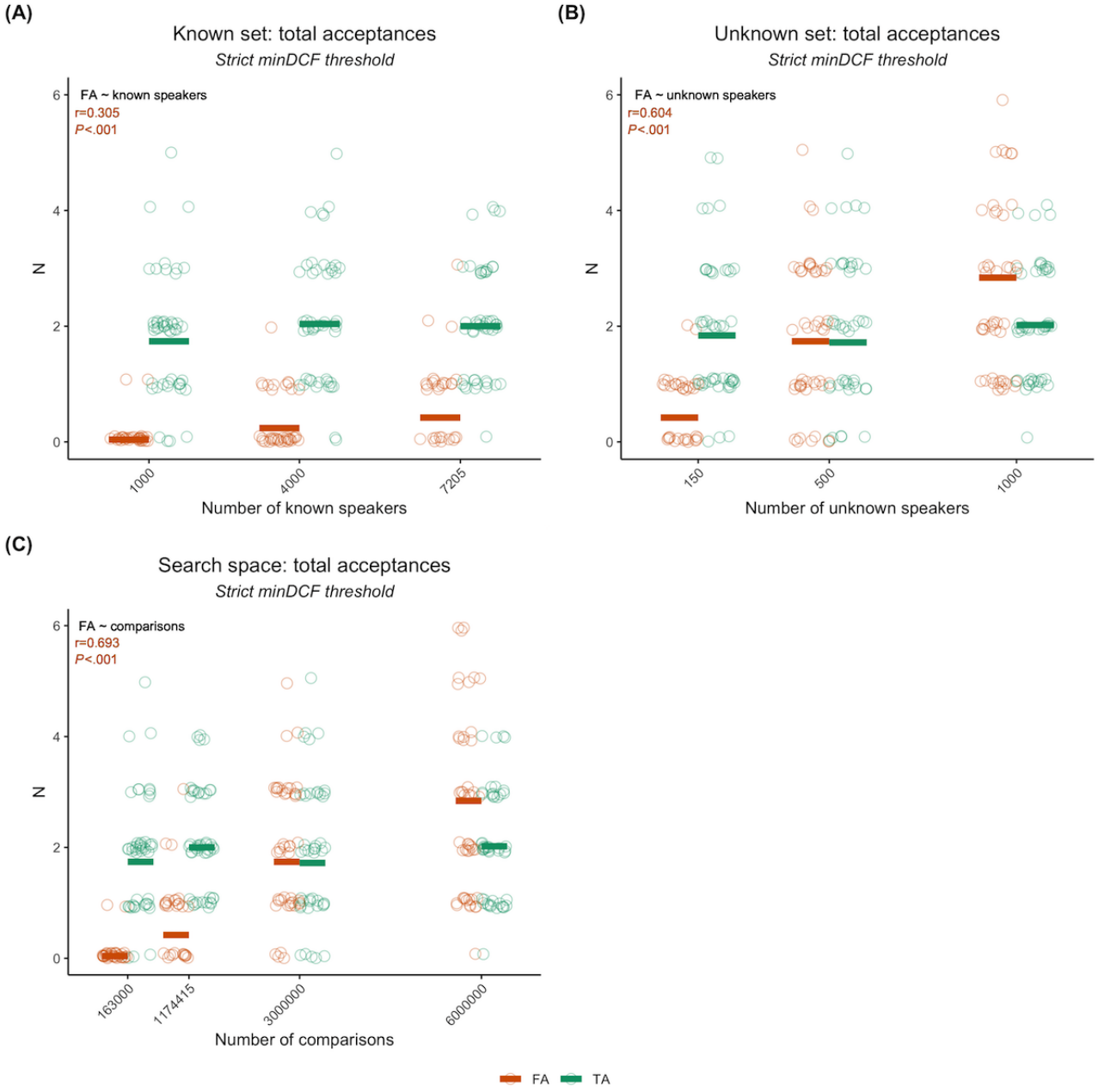
VoxCeleb Realistic Experiments: Effect of Search Space Size

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=known × 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.

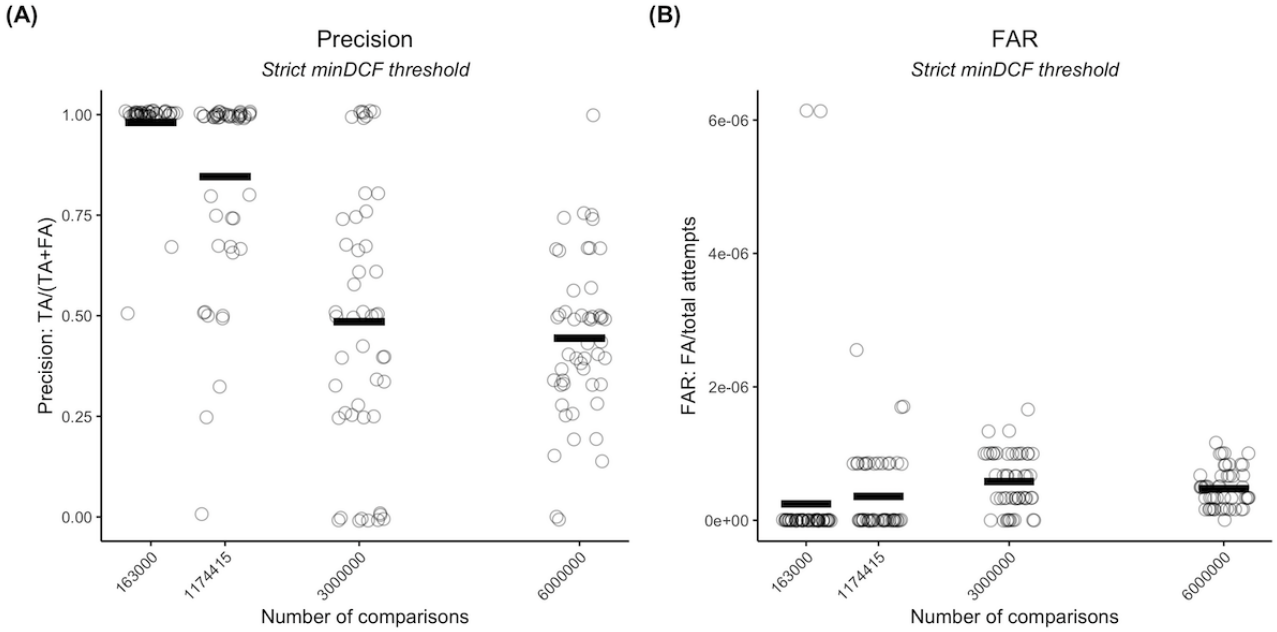


We observed that there was a high positive linear correlation between FAs and the number of comparisons (r=0.69; P<.001; t_{198}=13.54; Figure 2C), with the mean FAs increasing from 0.04 to 2.84 while TAs remained stable. The ratio between FA and TA (FA/TA) rose from 0.02 at 1×10^5 comparisons to 1.41 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.

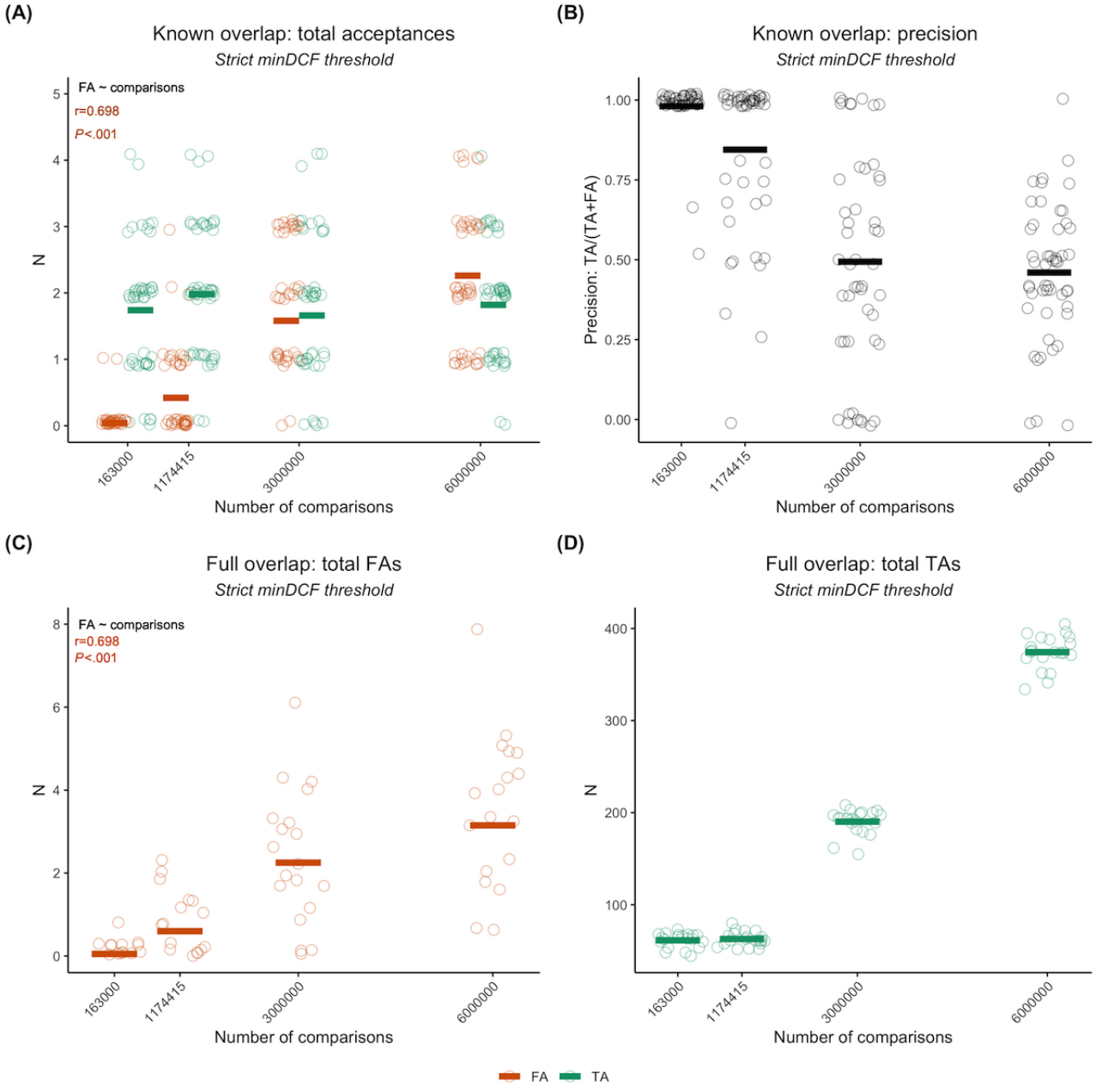


VoxCeleb Known-Overlap and Full-Overlap Experiments: Worst-Case Scenarios

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.



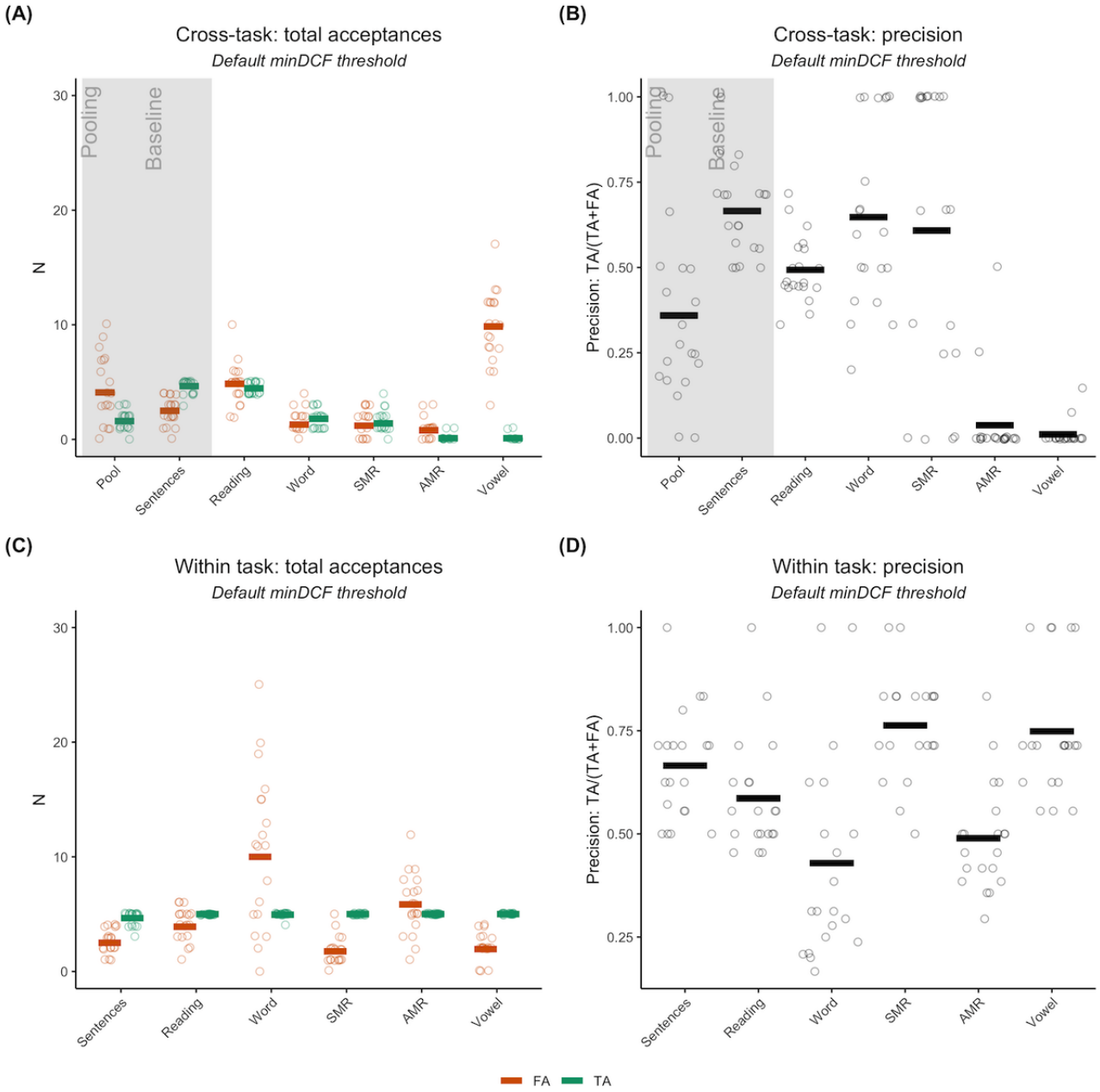
Mayo Clinic Speech Recording Experiments: Effect of Speech Task

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

IIn 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 21095825]
6. Pigoni 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 17213047]
13. Ribaric S, Pavešic 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]
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]
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] [Medline: 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] [Medline: 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] [Medline: 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]
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]
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]
24. Lu P, Zhu H, Sovernigo 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]
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]
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]
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] [Medline: 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]
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]
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]
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]
32. Desplanques B, Thienpondt J, Demuynck 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]
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]
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]
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]
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]
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]

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]
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]
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] [Medline: 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] [Medline: 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]
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]
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]
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]
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]
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]
52. Xia W, Liu Y, Wan Z, Vorobeychik Y, Kantacioglu 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] [Medline: 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] [Medline: 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]
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] [Medline: 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]
57. Rieke N, Hancox J, Li W, Milletarì 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] [Medline: 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]
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] [Medline: 33532283]
60. Xia W, Kantarcıoğlu 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]
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] [Medline: 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
PMID:38875581

Original Paper

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

Julian Späth[1], MSc; Zeno Sewald[2], BSc; Niklas Probul[1], MSc; Magali Berland[3], PhD; Mathieu Almeida[3], PhD; Nicolas Pons[3], PhD; Emmanuelle Le Chatelier[3], PhD; Pere Ginès[4,5,6,7], MD, PhD; Cristina Solé[4,5,6], MD; Adrià Juanola[4,5,6], MD, PhD; Josch Pauling[2], PhD; Jan Baumbach[1], Prof Dr

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

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

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

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

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

[6]Centro de Investigacion en Red de Enfermedades hepaticas y Digestivas (CIBEReHD), Madrid, Spain

[7]Faculty 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

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

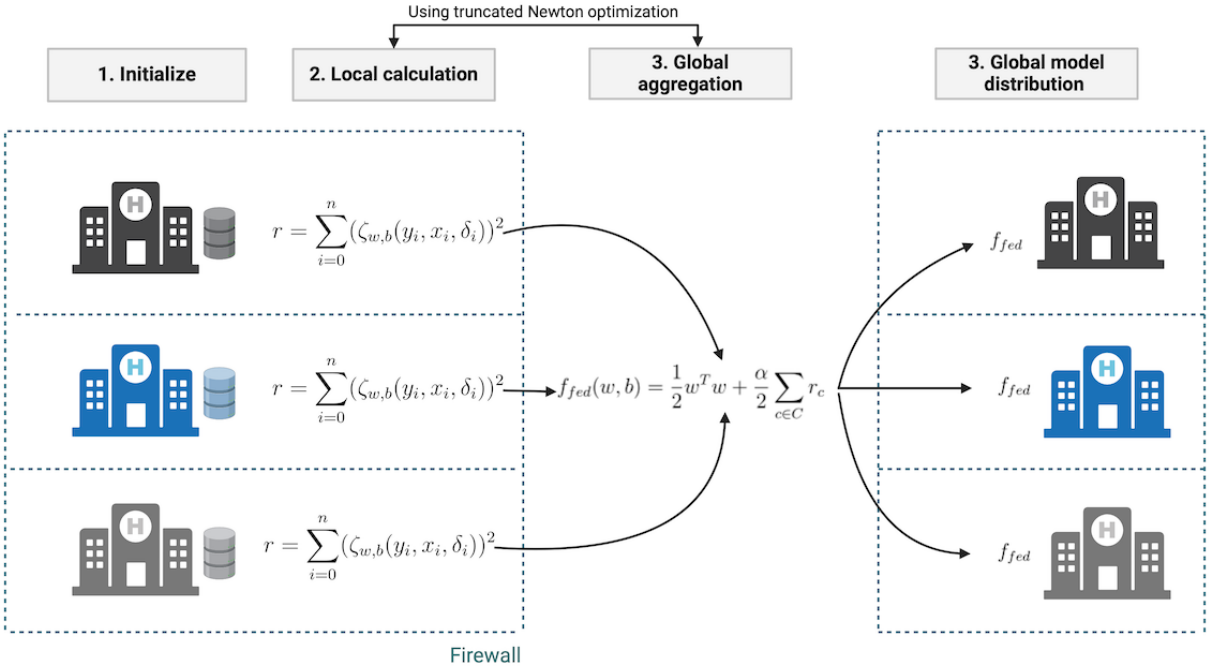
Here, we propose the adapted algorithm for the federated survival SVM, describe its implementation as a FeatureCloud app, and explain how we evaluated its performance.

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.

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.

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 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

[a]BRCA: breast cancer data set.

[b]GBSG2: German Breast Cancer Study Group 2 data set.

[c]WHAS500: 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.

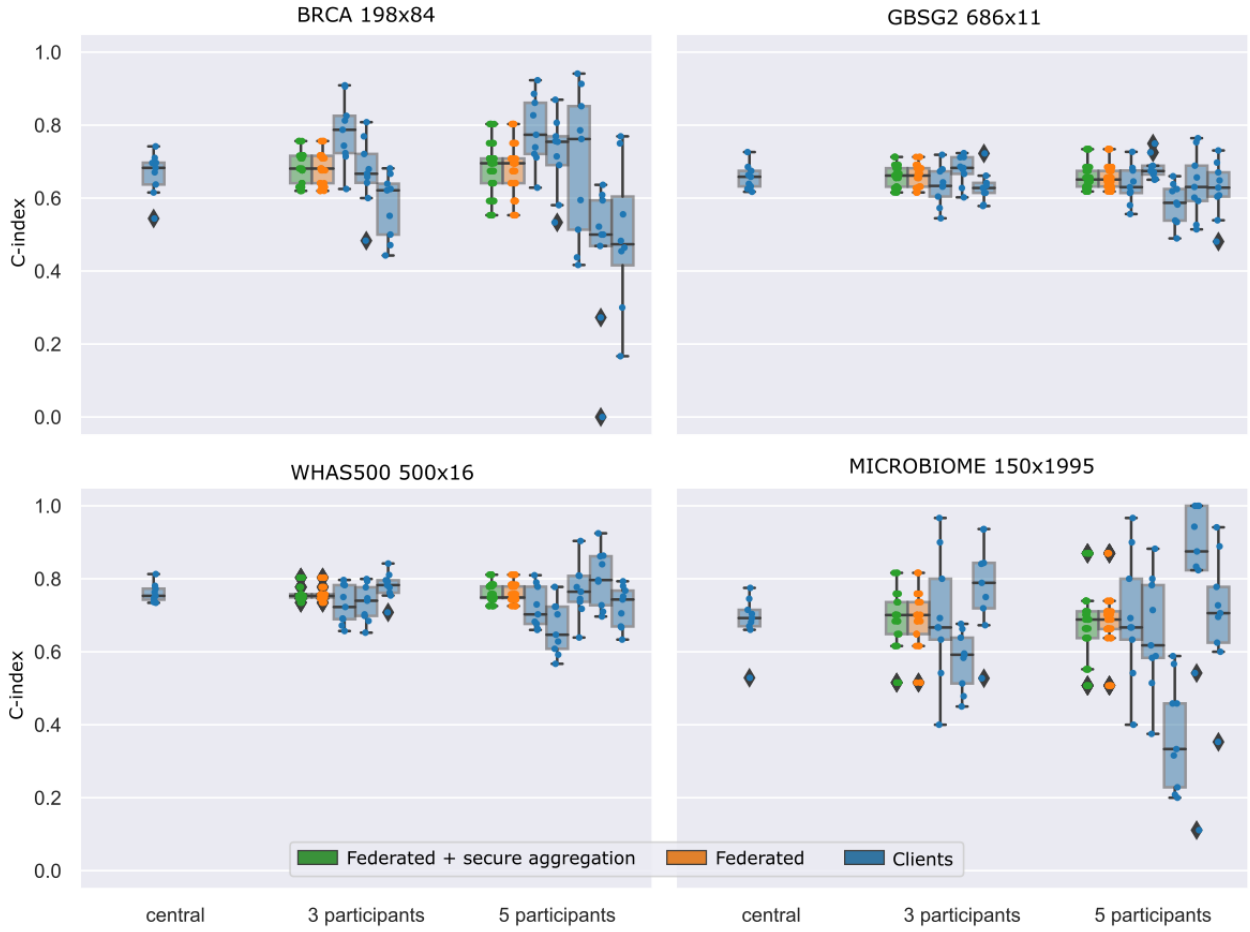
Our implementation of the federated survival SVM app uses all the functionalities offered by FeatureCloud and does not deviate from these best practices.

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 (GSBG2) 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 (GBSG2) 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.

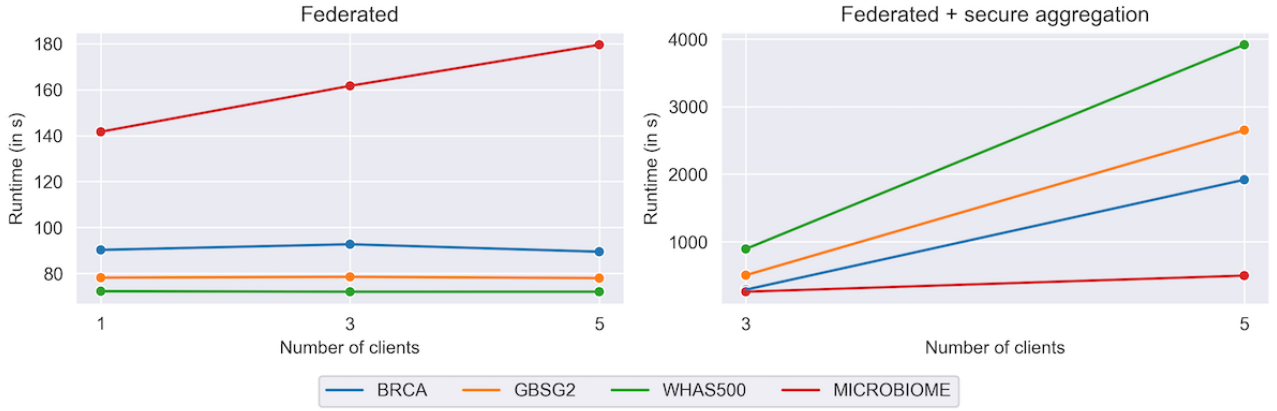
Our results further demonstrate the importance of large data sets, as the performance of the locally trained models on single clients (smaller sample size) shows a much higher variance than our federated models. If 5 institutes combine their small data sets, they can perform a much more reliable time-to-event 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 recieved speakig fees from Abbive 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]
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] [Medline: 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]
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]
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]
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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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]
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]
36. Hsu J, Gaboardi M, Haeberlen 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]
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] [Medline: 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]
41. Aalen O. Nonparametric inference for a family of counting processes. Ann Statist 1978;6(4):701-726. [doi: 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]
43. Ishwaran H, Kogalur UB, Blackstone EH, Lauer MS. Random survival forests. Ann Appl Stat 2008;2(3):841-860. [doi: 10.1214/08-aoas169]
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] [Medline: 35139148]
45. Shamir A. How to share a secret. Commun ACM 1979;22(11):612-613 [FREE Full text] [doi: 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]
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]
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
PMID:38875678

Original Paper

Online Health Search Via Multidimensional Information Quality Assessment Based on Deep Language Models: Algorithm Development and Validation

Boya Zhang[1], MSc; Nona Naderi[2], PhD; Rahul Mishra[1], PhD; Douglas Teodoro[1], PhD

[1]Department of Radiology and Medical Informatics, University of Geneva, Geneva, Switzerland

[2]Department 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

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 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.

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 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?".

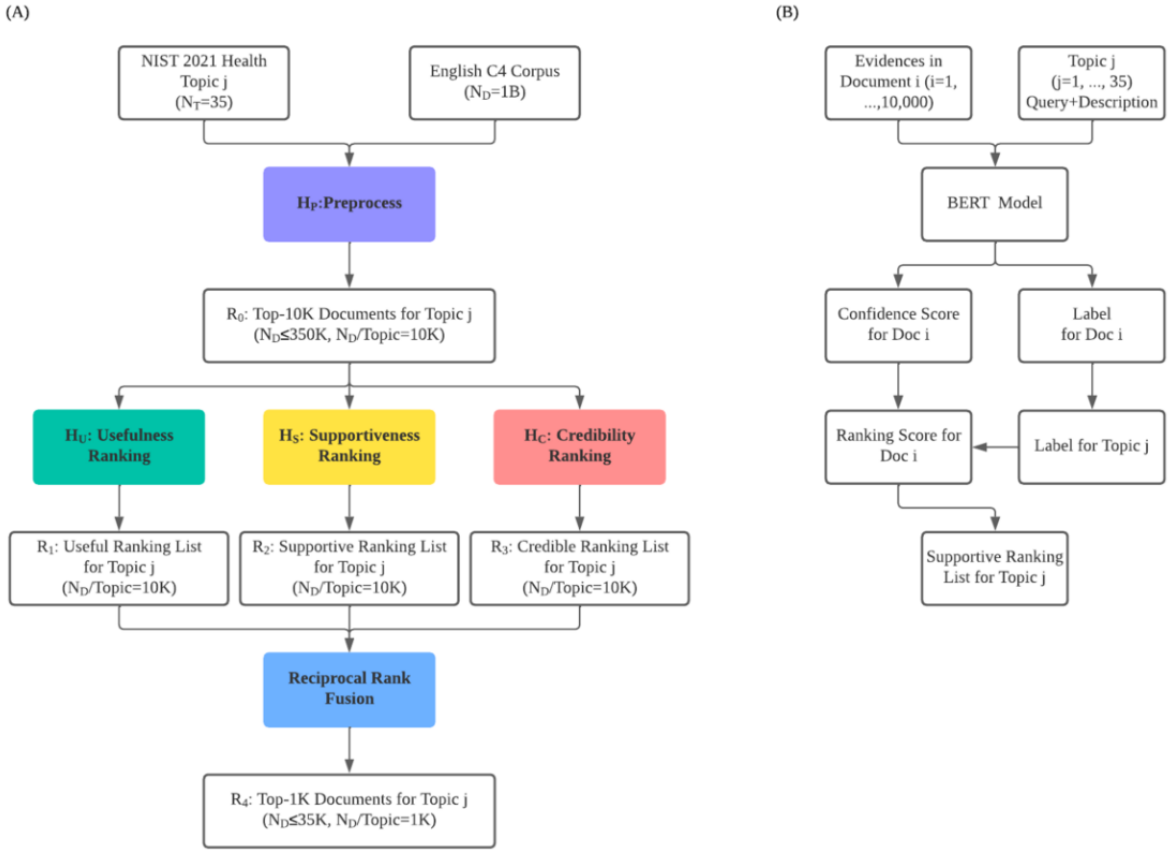
Article information	Article 1	Article 2
Doc ID	en.noclean.c4-train.07165-of-07168.96468	en.noclean.c4-train.00001-of-07168.126948
Time stamp	2019-04-25T18:00:17Z	2019-04-23T20:13:31Z
Text	[...] Do Not Use Sponging to Reduce a Fever. It is not recommended that you use sponging to reduce your child's fever. 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 adaption. 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 outcoming 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 implemations 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 aceess.

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 Multmedia 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

[a]nDCG: normalized discounted cumulative gain.

[b]Help: results considering only helpful documents in the relevance judgment.

[c]Harm: results considering only harmful documents in the relevance judgment.

[d]all_T: all topics.

[e]$helpful_T$: helpful topics.

[f]$unhelpful_T$: unhelpful topics.

[g]BM25: Best Match 25.

[h]Best performance.

[i]H_U: usefulness model.

[j]Best 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 adaption 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 adaption 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

[a]nDCG: normalized discounted cumulative gain.

[b]Help: results considering only helpful documents in the relevance judgment.

[c]Harm: results considering only harmful documents in the relevance judgment.

[d]all$_T$: all topics.

[e]helpful$_T$: helpful topics.

[f]unhelpful$_T$: unhelpful topics.

[g]H_p: preprocess.

[h]Best performance.

[i]H_U': usefulness model.

[j]Unlike 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%; *Hp*: P<.001; *Hu*': 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).

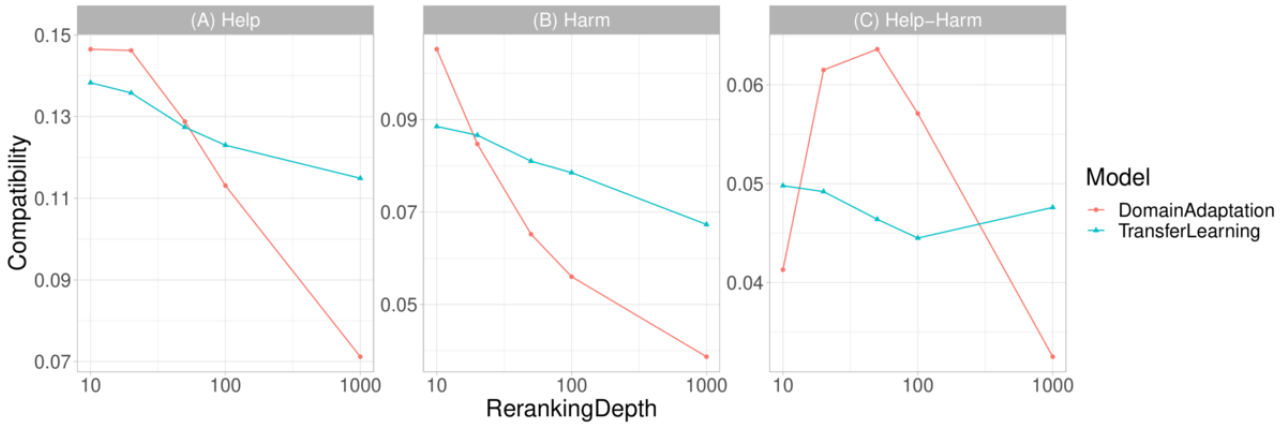
Reranking of the Top-N Documents

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.



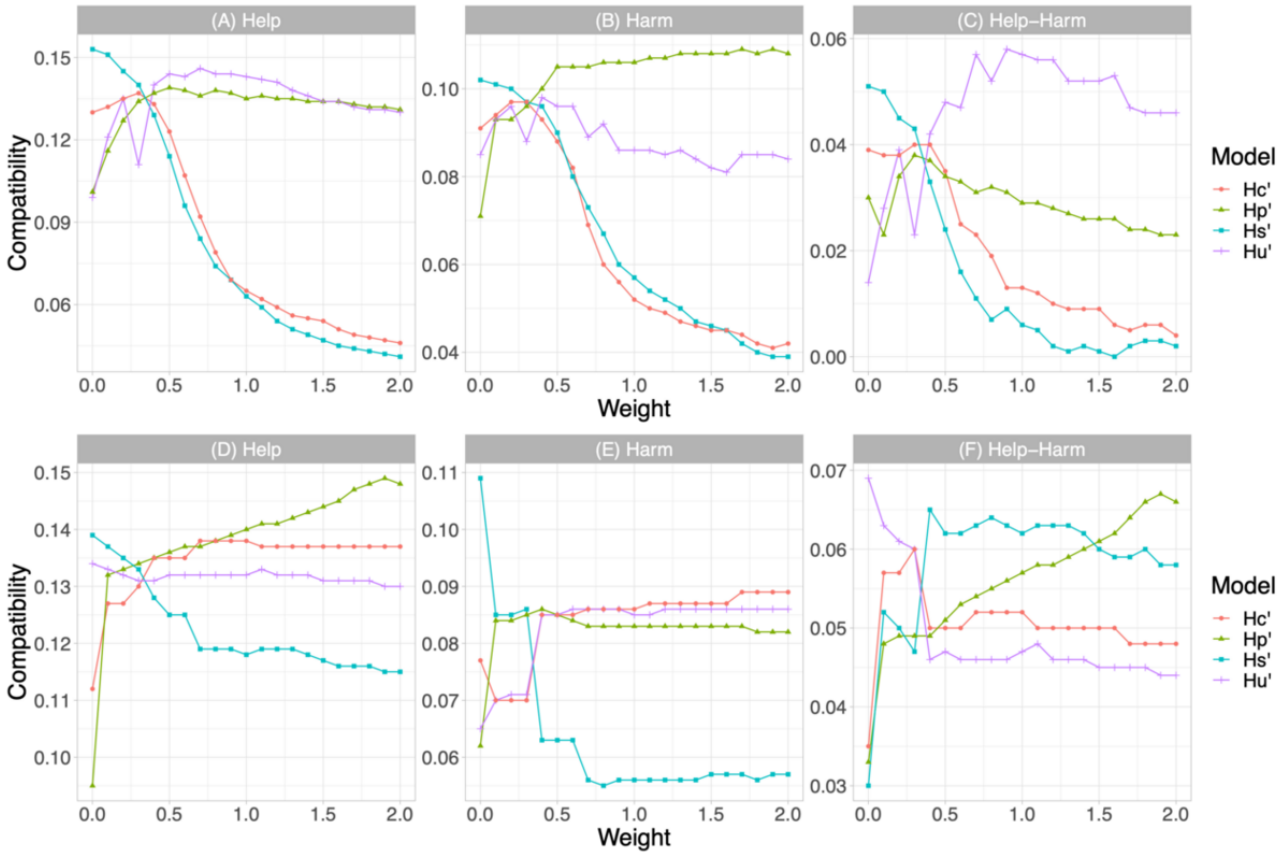
Quality Control

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

[a]Help: results considering only helpful documents in the relevance judgment.

[b]Harm: results considering only harmful documents in the relevance judgment.

[c]H_U: usefulness model.

[d]Best 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 funding 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] [Medline: 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] [Medline: 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] [Medline: 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]
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] [Medline: 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]
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] [Medline: 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]
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] [Medline: 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] [Medline: 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]
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, Goeuriot 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]
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]
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]
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]
28. Teodoro D, Ferdowsi S, Borissov 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] [Medline: 34375298]
29. Fernández-Pichel M, Losada DE, Pichel JC, Elsweiler D. Comparing TraditionalNeural 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]
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]
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]
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]
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 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]
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]
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] [Medline: 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] [Medline: 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]
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]
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]
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]
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]
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]
53. Stammbach 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]
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]
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, Rüger 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] [Medline: 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]
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]
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]
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]
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]
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]
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]
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] [Medline: 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
PMID:38875551

Original Paper

Identifying Patterns of Smoking Cessation App Feature Use That Predict Successful Quitting: Secondary Analysis of Experimental Data Leveraging Machine Learning

Leeann Nicole Siegel[1], MPH, PhD; Kara P Wiseman[2], MPH, PhD; Alex Budenz[1], MA, DrPH; Yvonne Prutzman[1], MPH, PhD

[1]National Cancer Instiute, National Institutes of Health, Rockville, MD, United States

[2]University 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

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. Hoeppner 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, Hoeppner 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 quitSTARTEMA 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)

[a]EMA: ecological momentary assessment.

[b]PHQ-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)
nbuttonsshared_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)

[a]SML: supervised machine learning.

[b]EMA: 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.

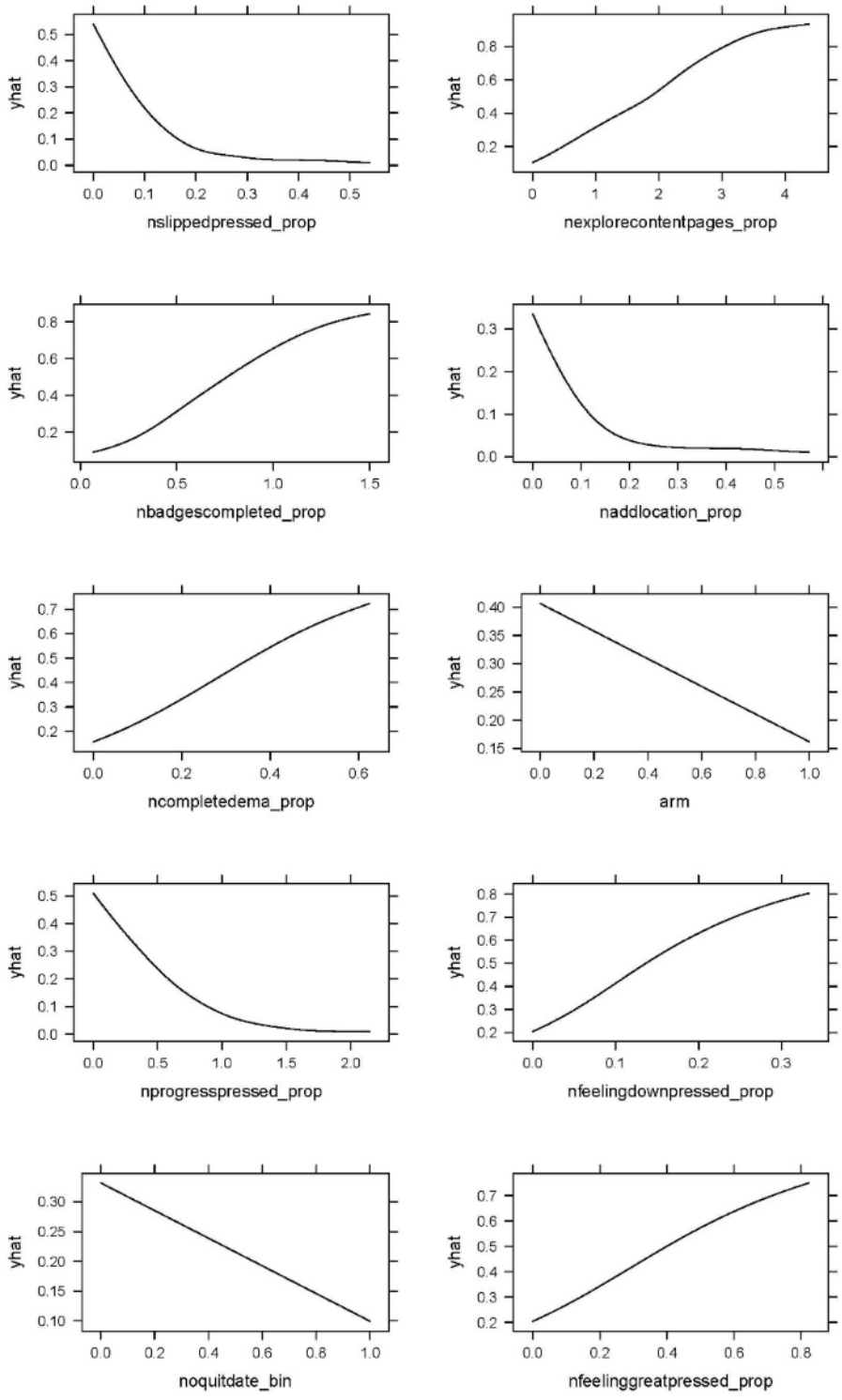
After building our SML model and assessing feature importance in the training set, we fit the model in the test set. The model's accuracy was 0.67, and both its sensitivity and specificity were also 0.67. We retained the SML model–predicted probabilities of cessation as a variable in the test set.

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 α=.05 level (χ^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 (χ^2_{10}=7.0; *P*=.73). The likelihood ratio test comparing model 2 to model 1 was not statistically significant (χ^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 (χ^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 on smoking cessation apps and health behavior apps more broadly to inform the design and refinement of such apps.

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] [Medline: 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] [Medline: 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, Laracuente 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]
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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 34347865]
13. Hoeppner BB, Hoeppner 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 25397860]
22. Hoepper 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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]
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]
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] [Medline: 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
PMID:38875564

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 Nohara[4], PhD; Yoshinobu Wakisaka[2], MD, PhD; Tetsuro Ago[2,5], MD, PhD; Naoki Nakashima[6], MD, PhD; Takanari Kitazono[2,5], MD, PhD; Masahiro Kamouchi[1,5], MD, PhD[‡]

[1]Department of Health Care Administration and Management, Graduate School of Medical Sciences, Kyushu University, Fukuoka, Japan

[2]Department of Medicine and Clinical Science, Graduate School of Medical Sciences, Kyushu University, Fukuoka, Japan

[3]Biostatistics Center, Graduate School of Medicine, Kurume University, Kurume, Japan

[4]Big Data Science and Technology, Faculty of Advanced Science and Technology, Kumamoto University, Kumamoto, Japan

[5]Center for Cohort Studies, Graduate School of Medical Sciences, Kyushu University, Fukuoka, Japan

[6]Medical 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

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

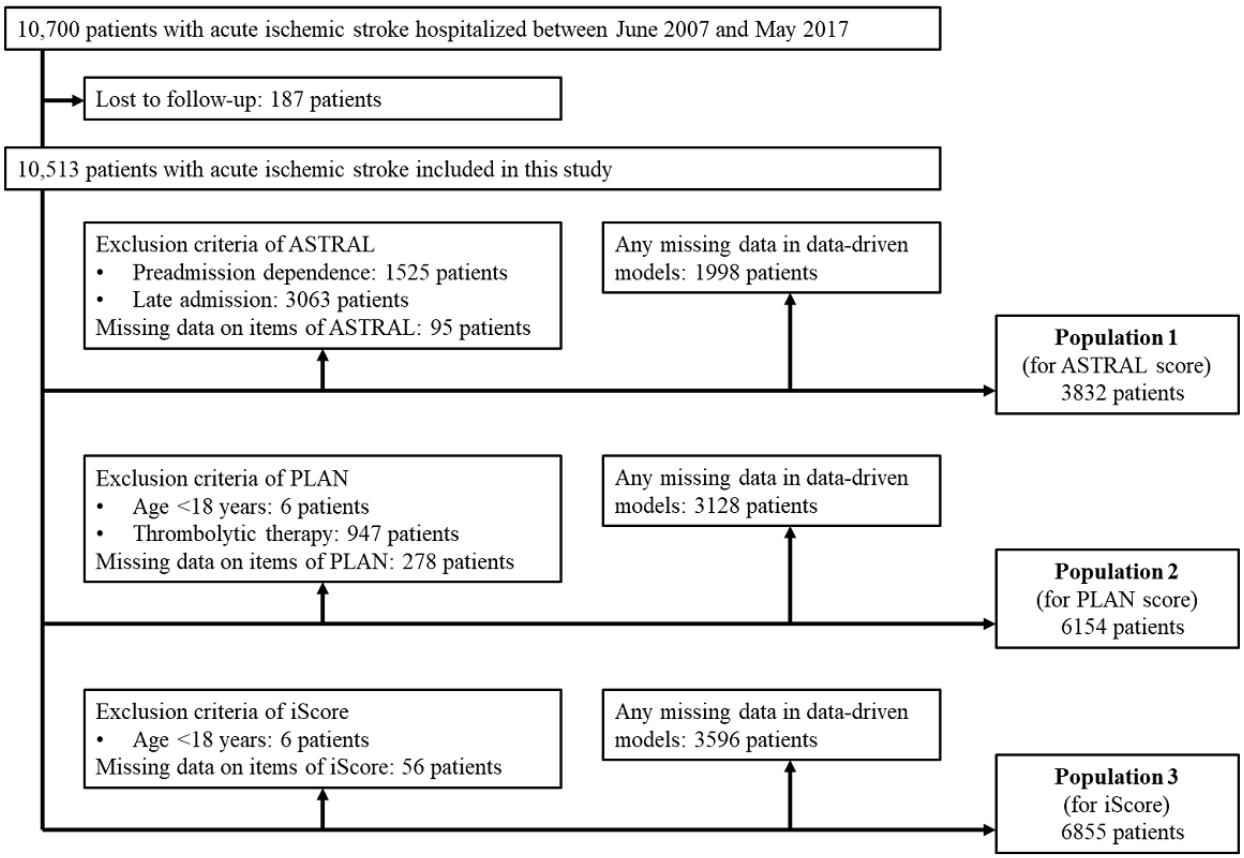
The conventional stroke prognostic scores were used for comparison against data-driven prediction models. In this study, we selected prognostic scores based on the following criteria: they are multiitem and point-based scores using demographic and clinical information, they were developed to predict 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

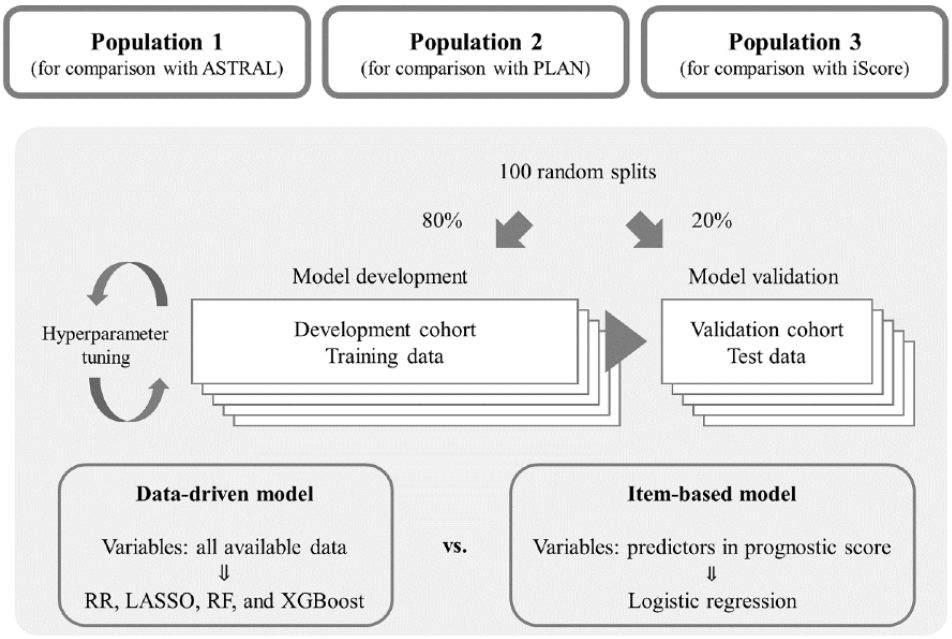
The discriminative ability of each model was evaluated using the area under the receiver operating characteristic curve (AUROC) and the area under the precision-recall curve (AUPRC). AUPRC was calculated because it is a useful performance metric for unbalanced data of infrequent outcome events, such 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 $1/N \sum^{N}_{i=1} (pi–ai)^2$, $(0 \leq BS \leq 1)$, where pi is the predicted probability of the occurrence of an event ranging from 0 to 1, ai 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 (%)				<.001
≤1 h	943 (9)	490 (7.7)	453 (11)	
≤3 h	1469 (14)	771 (12)	698 (17)	
≤6 h	1141 (10.9)	644 (10.1)	497 (12.1)	
≤24 h	3515 (33.4)	2090 (32.6)	1425 (34.7)	
>24 h	3445 (32.8)	2410 (37.6)	1035 (25.2)	
Stroke subtype, n (%)				<.001
Small vessel occlusion	2119 (20.2)	1724 (26.9)	395 (9.6)	
Large artery atherosclerosis	1823 (17.3)	1006 (15.7)	817 (19.9)	
Cardioembolism	2496 (23.7)	1054 (16.5)	1442 (35.1)	
Other determined etiology	2146 (20.4)	1404 (21.9)	742 (18.1)	
Undetermined	1929 (18.3)	1217 (19)	712 (17.3)	
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

[a]mRS: modified Rankin Scale.
[b]NIHSS: National Institutes of Health Stroke Scale.
[c]SBP: systolic blood pressure.
[d]DBP: diastolic blood pressure.
[e]WBC: white blood cell count.
[f]RBC: red blood cell count.
[g]AST: aspartate aminotransferase.
[h]ALT: alanine aminotransferase.
[i]LDH: lactate dehydrogenase.
[j]ALP: alkaline phosphatase.
[k]BUN: blood urea nitrogen.
[l]eGFR: estimated glomerular filtration rate.
[m]hsCRP: high-sensitivity C-reactive protein.
[n]PT-INR: international normalized ratio of prothrombin time.
[o]APTT: activated partial thromboplastin time.

Assessment of Model Performance

AUROCs 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)

[a]The 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).

[b]RR: ridge regression.

[c]LASSO: least absolute shrinkage and selection operator regression.

[d]RF: random forest.

[e]XGBoost: 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)

[a]The 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).

[b]RR: ridge regression.

[c]LASSO: least absolute shrinkage and selection operator regression.

[d]RF: random forest.

[e]XGBoost: 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)

[a]The 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).

[b]RR: ridge regression.

[c]LASSO: least absolute shrinkage and selection operator regression.

[d]RF: random forest.

[e]XGBoost: 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
Bier 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

[a]Data 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.

[b]RR: ridge regression.

[c]LASSO: least absolute shrinkage and selection operator regression.

[d]RF: random forest.

[e]XGBoost: Extreme Gradient Boosting.

[f]AUROC: area under the receiver operating characteristic curve.

[g]AUPRC: 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.

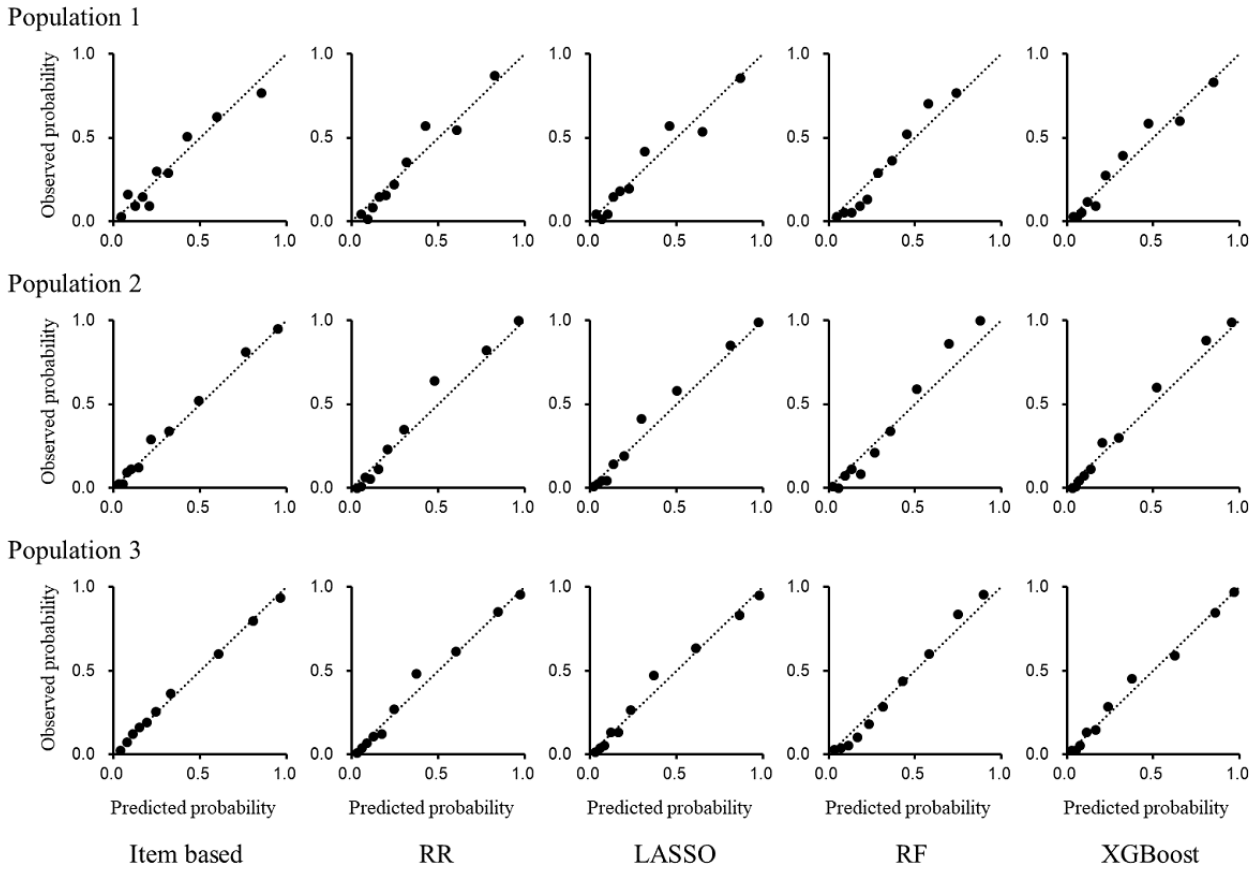
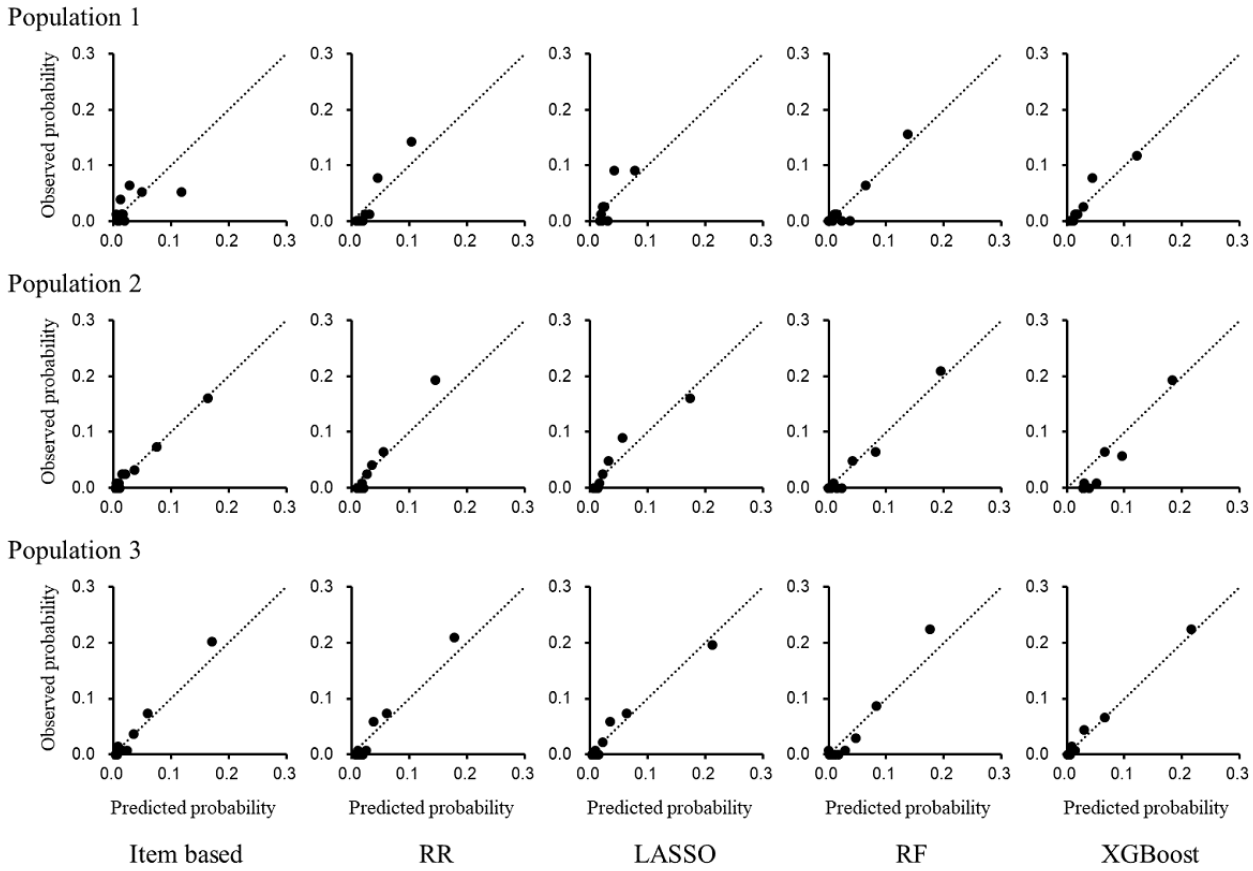


Figure 4. Calibration of item-based and data-driven models for predicting death. Calibration for predicting death was compared between the item-based regression model and data-driven models (ridge regression [RR], least absolute shrinkage and selection operator regression [LASSO], random forest [RF], 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.

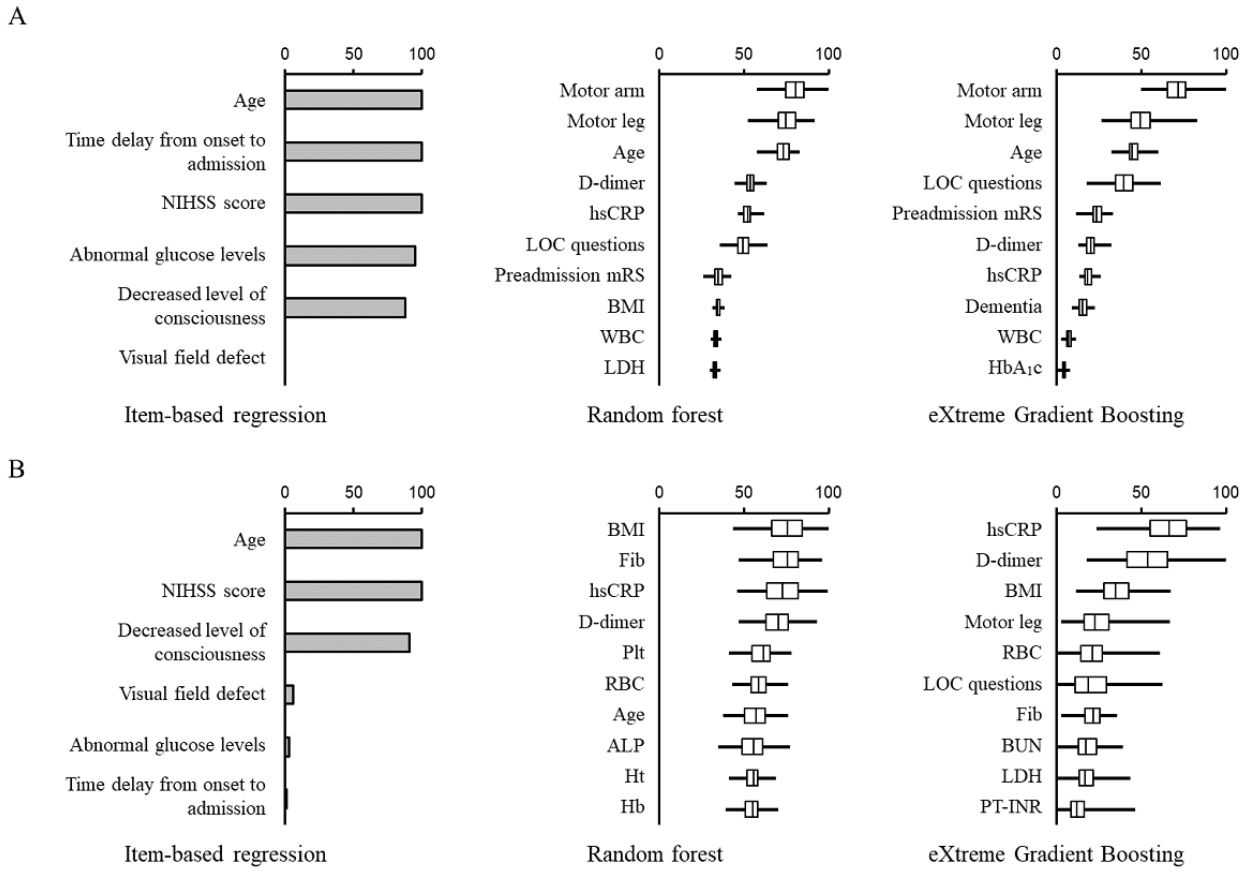


Figure 6. Comparison of variable importance between items of the preadmission comorbidities, level of consciousness, age, and neurologic deficit (PLAN) score and explanatory variables in machine learning model in population 2. The contribution of each variable to the models in predicting poor functional 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.

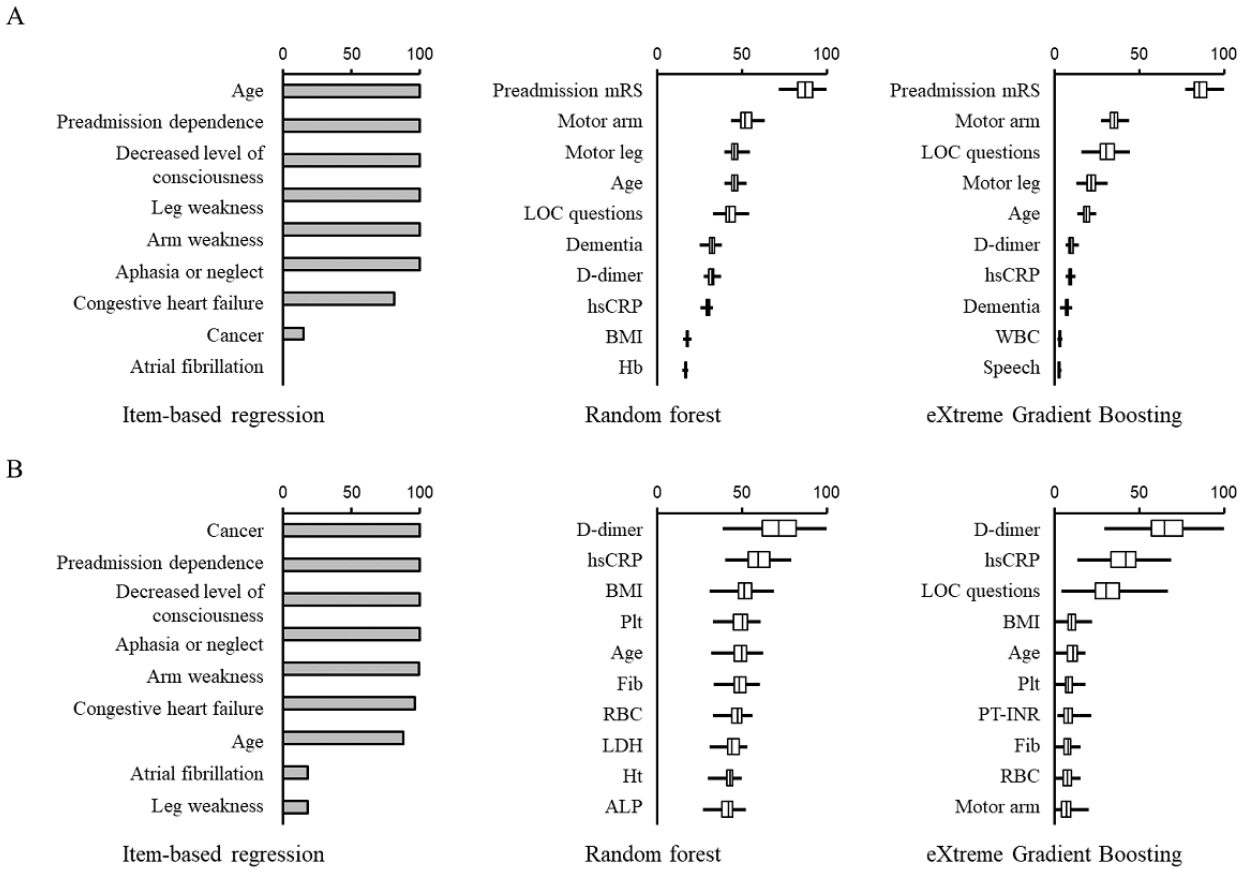


Figure 7. Comparison of variable importance between items of Ischemic Stroke Predictive Risk Score (iScore) and explanatory variables in machine learning model in population 3. The contribution of each variable to the models in predicting poor functional outcome (A) and death (B) is shown. The patients 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.

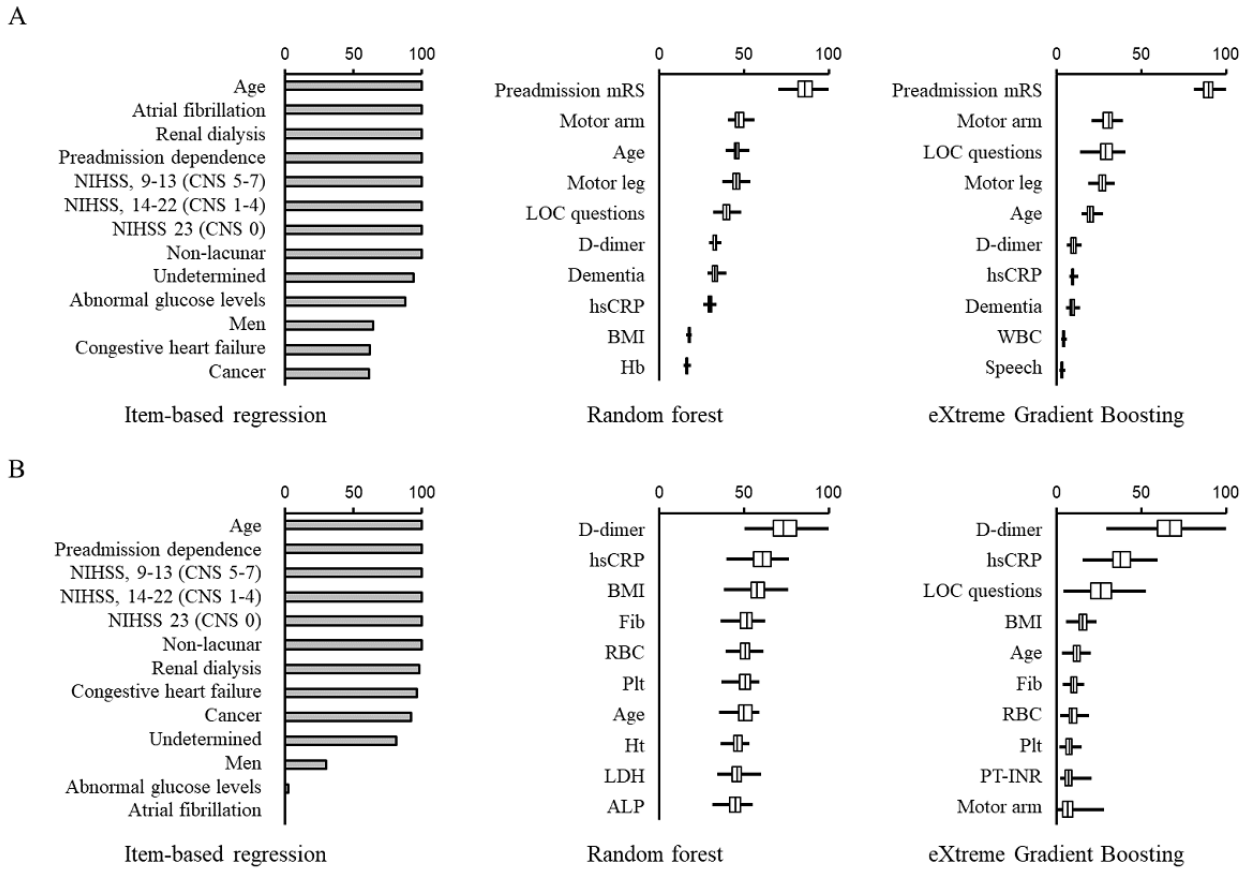
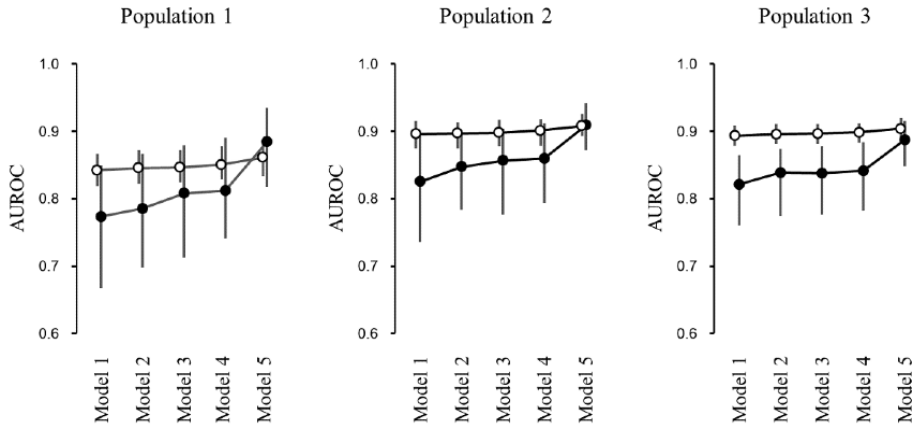


Figure 8. Improvement of discrimination in a data-driven model by adding different types of data. The area under the receiver operating characteristic curves (AUROCs) for predicting poor functional outcome (open circles) and death (closed circles) were compared among the 5 models, which used different 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 (Hakujyuji 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.0000000000000158] [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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 30763612]
26. Gravesteijn 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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]
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]
33. Breiman L. Random forests. Mach Learn 2001;45:5-32. [doi: 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]
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] [Medline: 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]
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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 33412918]
43. Phipps MS, Cronin CA. Management of acute ischemic stroke. BMJ 2020 Feb 13;368:l6983. [doi: 10.1136/bmj.l6983] [Medline: 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] [Medline: 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
PMID:38875590

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 Cheung[1], BSc; Kalya M Kee[1,4], BA, MSc; May O Lwin[4], BA, MBA, PhD; Peter J Schulz[1,4], BA, MA, PhD; Minhu Chen[5], MBBS, PhD; Kaichun Wu[6], MD, PhD; Simon SM Ng[7], MBChB, MD; Rashid Lui[8], MBChB; Tiing Leong Ang[9], MBBS, MRCPUK; Khay Guan Yeoh[10,11], MBBS, MMed; Han-mo Chiu[12,13], MD, PhD; Deng-chyang Wu[14], MD, PhD; Joseph JY Sung[1], MD, PhD

[1]Lee Kong Chian School of Medicine, Nanyang Technological University, Singapore, Singapore, Singapore

[2]School of Biological Sciences, Nanyang Technological University, Singapore, Singapore

[3]Center for Biomedical Informatics, Nanyang Technological University, Singapore, Singapore

[4]Wee Kim Wee School of Communication and Information, Nanyang Technological University, Singapore, Singapore

[5]The First Affiliated Hospital, Sun Yat-sen University, Guangzhou, China

[6]Xijing Hospital, Fourth Military Medical University, Xi'an, China

[7]Department of Surgery, Faculty of Medicine, The Chinese University of Hong Kong, Hong Kong, China (Hong Kong)

[8]Prince of Wales Hospital, Hospital Authority, Hong Kong, China (Hong Kong)

[9]Department of Gastroenterology and Hepatology, Changi General Hospital, SingHealth, Singapore, Singapore

[10]Department of Gastroenterology and Hepatology, National University Hospital, National University Health System, Singapore, Singapore

[11]Department of Medicine, Yong Loo Lin School of Medicine, National University of Singapore, Singapore, Singapore

[12]Department of Internal Medicine, National Taiwan University Hospital, Taiwan, China

[13]Department of Internal Medicine, College of Medicine, National Taiwan University, Taiwan, China

[14]Kaohsiung 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

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.

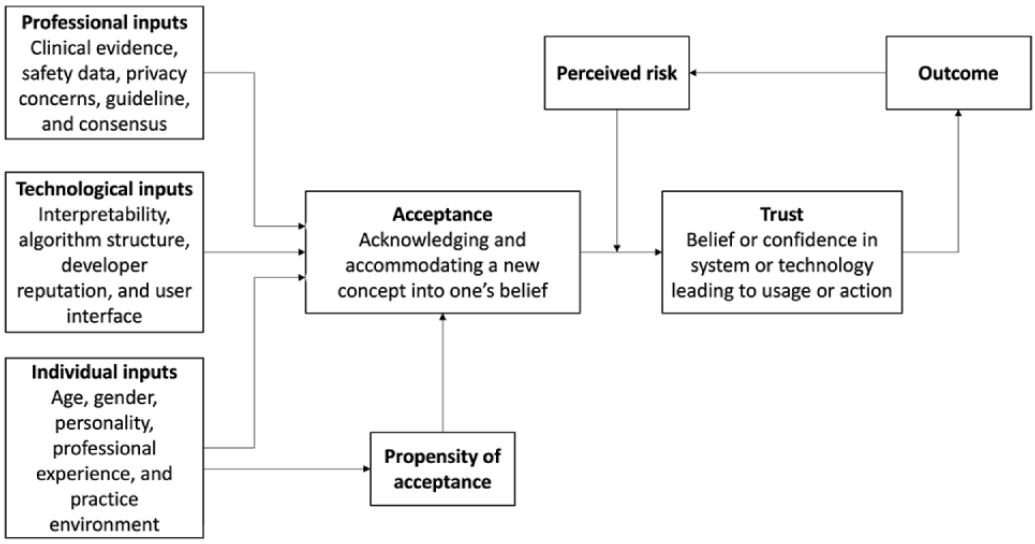
Scenario	Objective
Computer-aided detection: use of AI to assist in identifying the presence of colorectal polyps and improving adenoma detection rate.	To evaluate the acceptability of AI to assist in the interpretation of medical imagery in detecting colorectal lesions under different bowel preparations 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.

[a]AI: 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 α=.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)

[a]11 participants did not report their ages or gender.

[b]1 participant did not report their country or area.

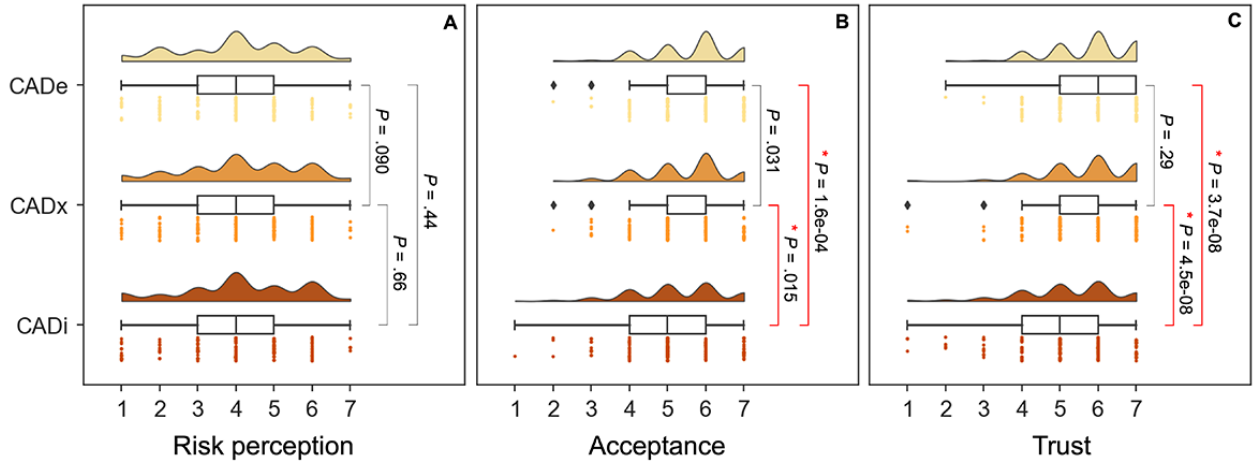
[c]3 participants did not report their main work setting, specialty, and years practicing in a specialty.

[d]4 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: $Median_{CADe}=Median_{CADx}=Median_{CADi}=4.0$; $Wilcox_{CADe\text{-}CADx}$: $P=.09$; $Wilcox_{CADe\text{-}CADi}$: $P=.44$; $Wilcox_{CADx\text{-}CADi}$: $P=.66$). 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: $Median_{CADe}=6.0$, $Median_{CADx}=6.0$, $Median_{CADi}=5.0$; $Wilcox_{CADe\text{-}CADx}$: $P=.031$; $Wilcox_{CADe\text{-}CADi}$: $P=1.6\times10^{-4}$; $Wilcox_{CADx\text{-}CADi}$: $P=.02$). Similarly for trust, CADe and CADx were rated higher than CADi (Figure 2C: $Median_{CADe}=6.0$, $Median_{CADx}=6.0$, $Median_{CADi}=5.0$; $Wilcox_{CADe\text{-}CADx}$: $P=.29$; $Wilcox_{CADe\text{-}CADi}$: $P=3.7\times10^{-08}$; $Wilcox_{CADx\text{-}CADi}$: $P=4.5\times10^{-08}$).

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 Wilcox 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 ≤.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: $P=9.7\times10^{-6}$; CADx: $P=1.7\times10^{-06}$; CADi: $P=3.3\times10^{-04}$). 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.

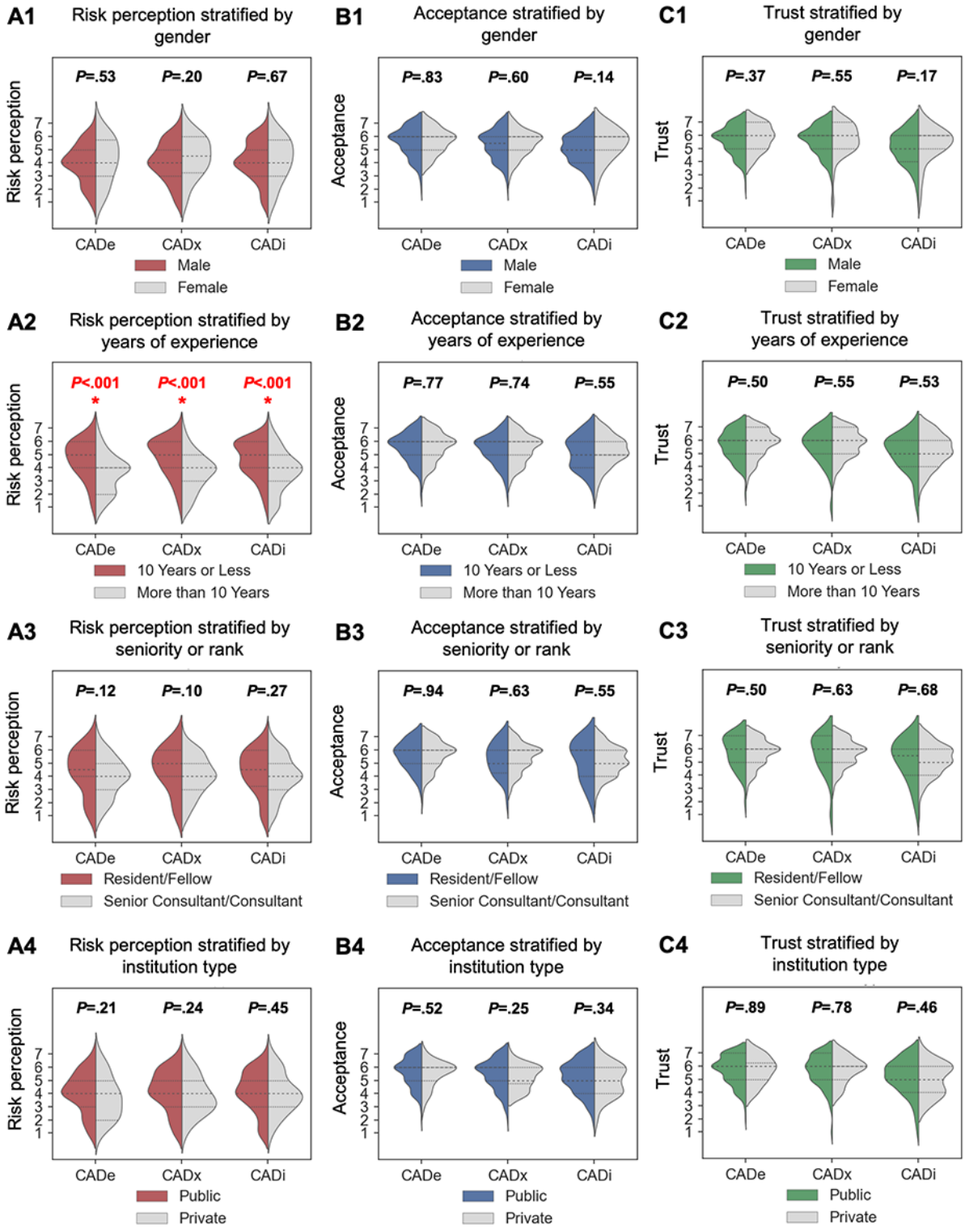
Finally, we compared practitioners from public hospitals with those from private hospitals (Figure 3A4, 3B4, and 3C4). There was no statistically significant difference between private hospital practitioners against their public counterparts, although there was a noticeable difference in CADx on acceptance (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=1.6×10^{-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 ≤.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_app1.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] [Medline: 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] [Medline: 35576007]
3. Hoiland RL, Rikhraj 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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]
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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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]
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]
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] [Medline: 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] [Medline: 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]
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]
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]
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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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]
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]

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
PMID:38875591

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 Ewals[1], MSc; Lynn J J Heesterbeek[2], MSc; Bin Yu[3], PhD; Kasper van der Wulp[1], MD; Dimitrios Mavroeidis[4], PhD; Mathias Funk[5], PhD; Chris C P Snijders[6], Prof Dr; Igor Jacobs[7], PhD; Joost Nederend[1], MD, PhD; Jon R Pluyter[2], PhD; e/MTIC Oncology group[8]

[1]Catharina Cancer Institute, Catharina Hospital Eindhoven, Eindhoven, Netherlands
[2]Department of Experience Design, Royal Philips, Eindhoven, Netherlands
[3]Research Center for Marketing and Supply Chain Management, Nyenrode Business University, Breukelen, Netherlands
[4]Department of Data Science, Philips Research, Eindhoven, Netherlands
[5]Department of Industrial Design, Eindhoven University of Technology, Eindhoven, Netherlands
[6]Department of Human Technology Interaction, Eindhoven University of Technology, Eindhoven, Netherlands
[7]Department of Hospital Services and Informatics, Philips Research, Eindhoven, Netherlands
[8]See 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

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

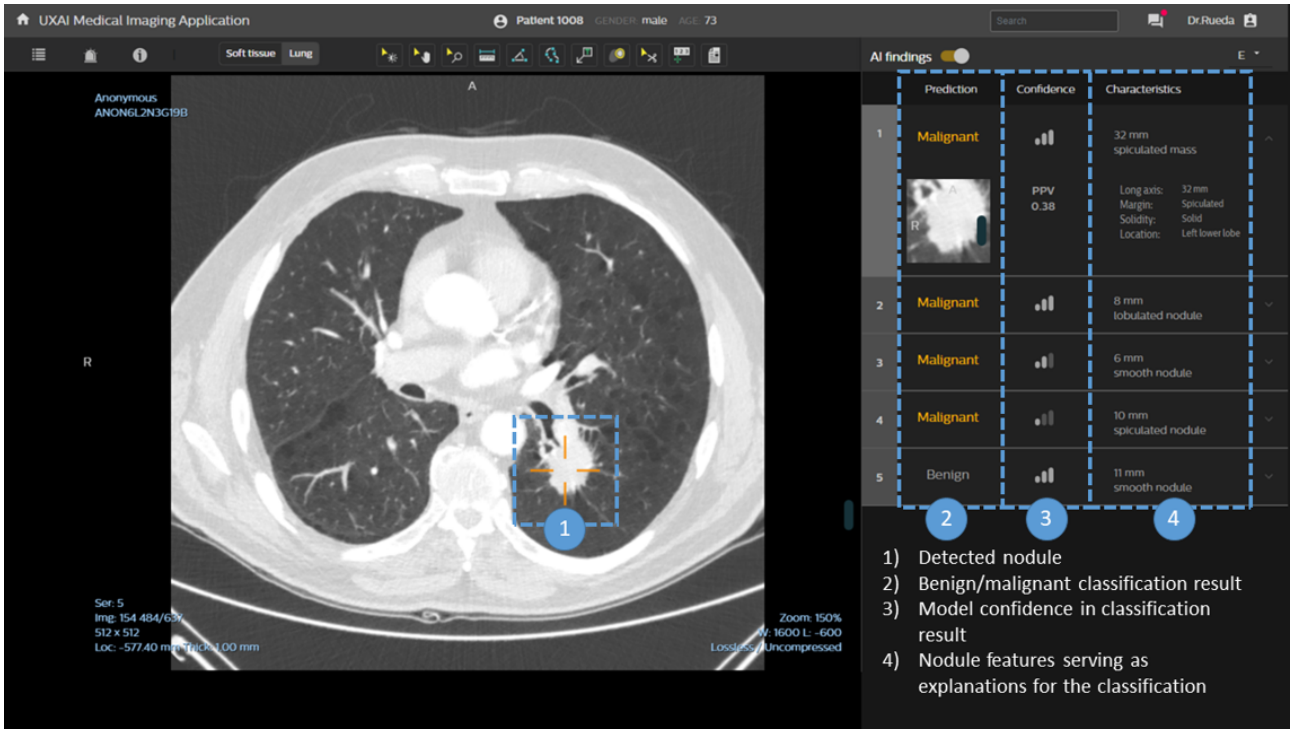
We tested the hypotheses using a 2×2 repeated-measures quasi-experimental design: informative versus reflective onboarding tutorial and black box versus explainable AI output. In this simulated use study, we aimed to realistically mimic clinical practice [33,34]. Realistic clinical simulations allow 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, brainstorms, 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.

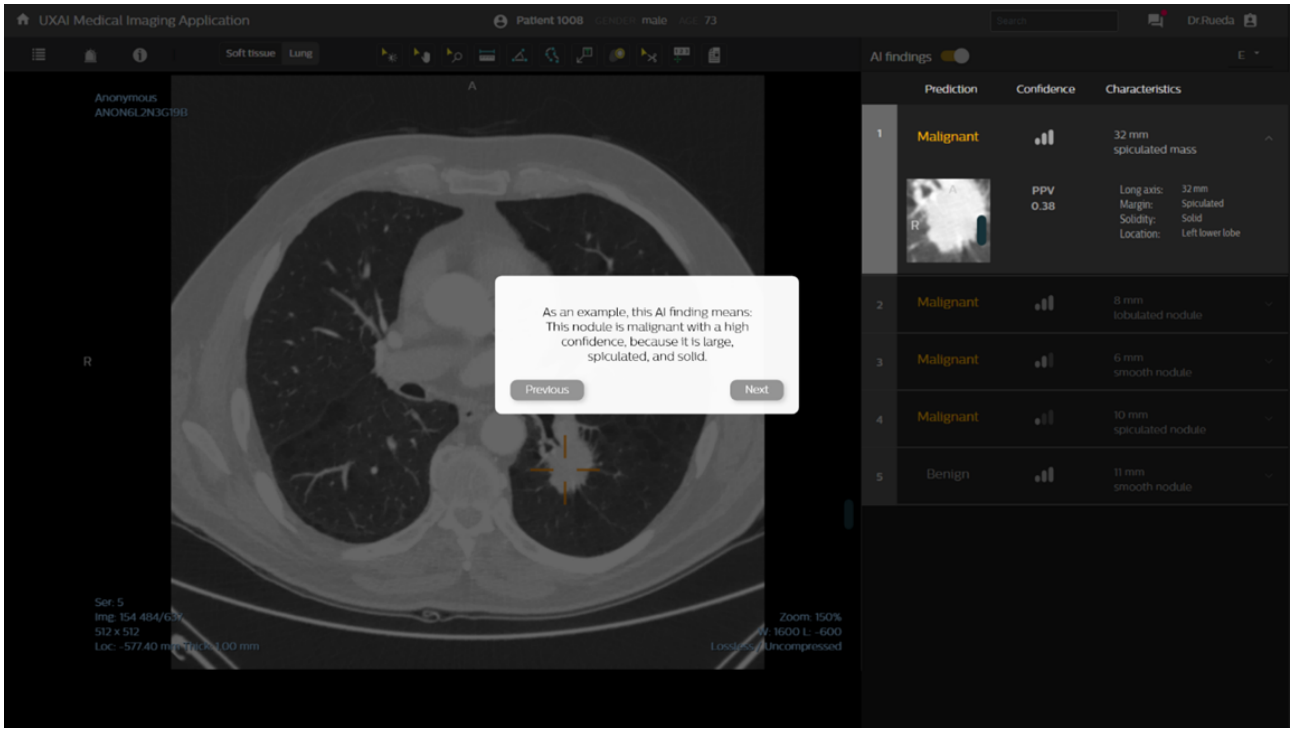
Confidence in the malignancy classification (cue 3) was given by means of PPVs for malignant predictions, indicating the probability that nodules with malignant predictions were actually malignant, and by means of NPVs for benign predictions, indicating the probability that nodules with a benign prediction 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.

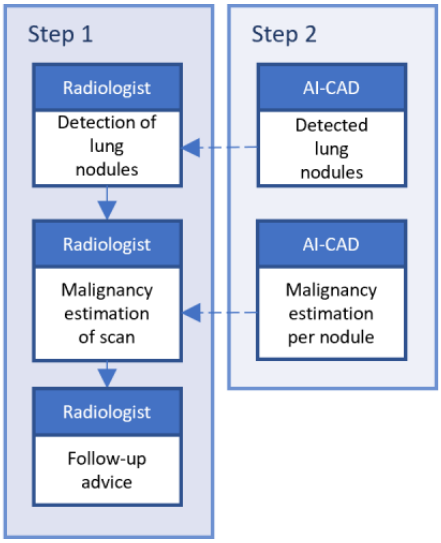


Study Protocol

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 2×2 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.

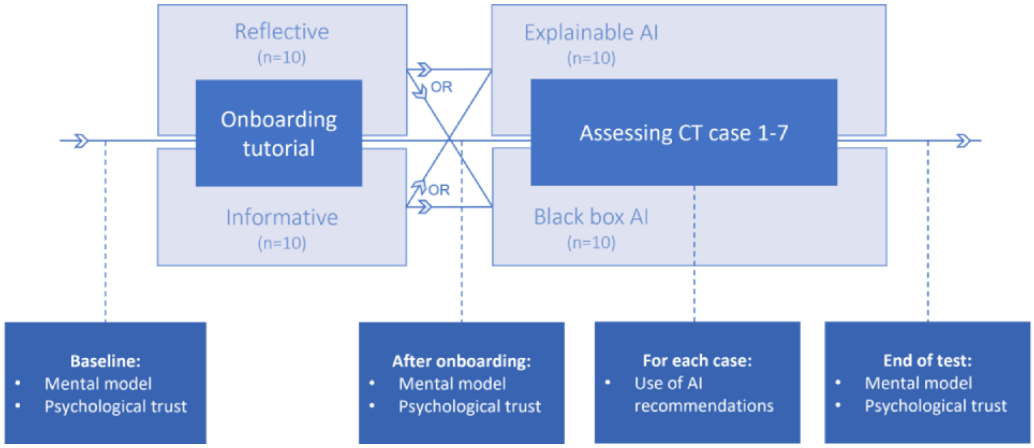
Figure 3. Flow diagram showing the clinical decisions of radiologists, which might potentially be influenced by the outcomes of the artificial intelligence model. The detected nodules may influence the malignancy estimation, and the malignancy estimation may influence the follow-up advice. AI-CAD: artificial 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

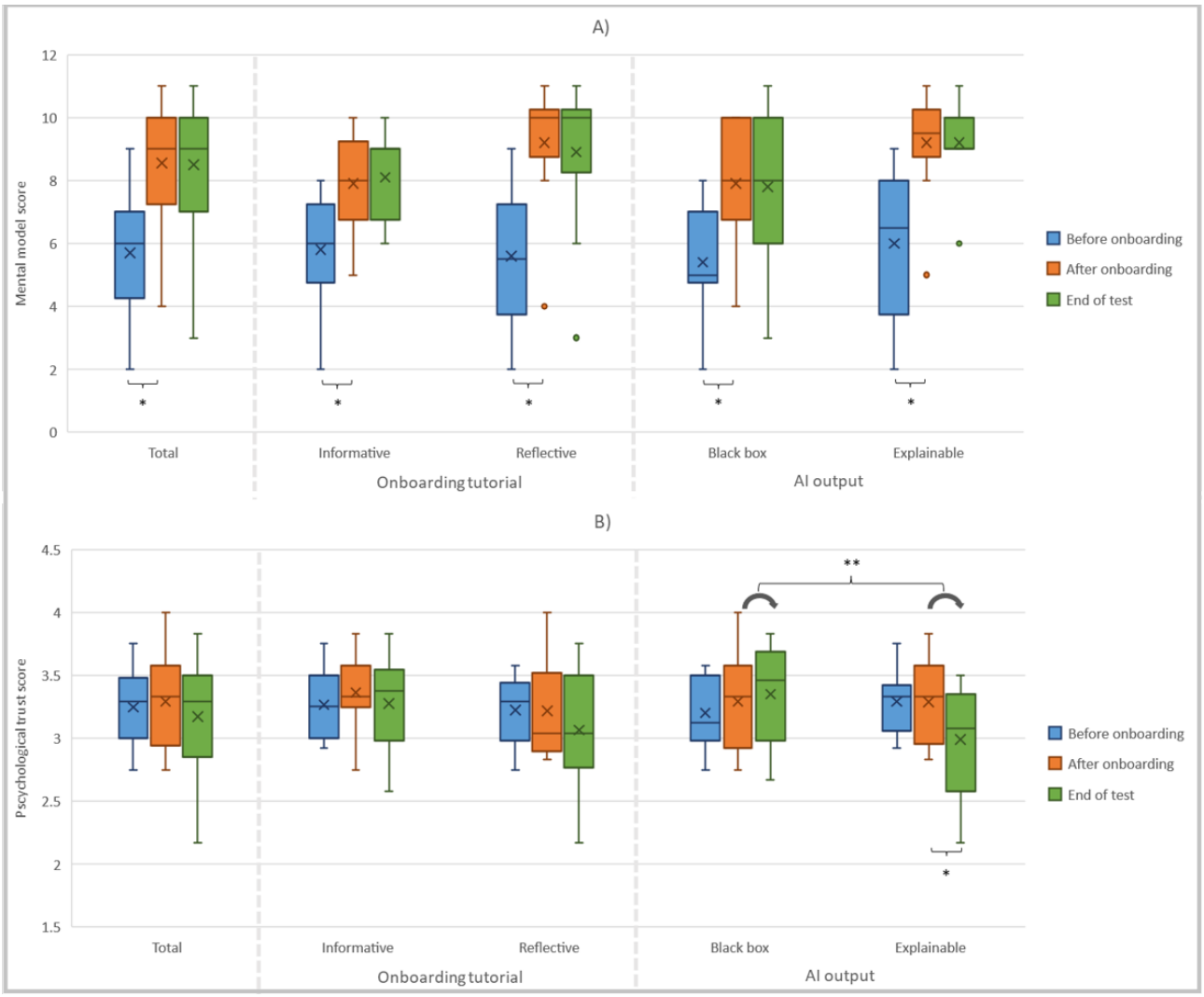
Figure 5 presents the mental model and psychological trust scores before onboarding, after onboarding, and at the end of the test. These scores were shown for all radiologists together and for the 2 variations of the onboarding tutorials and AI output separately.

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: β=0.143; *P*=.16; second model: β=0.167; *P*=.04; third model: β=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.

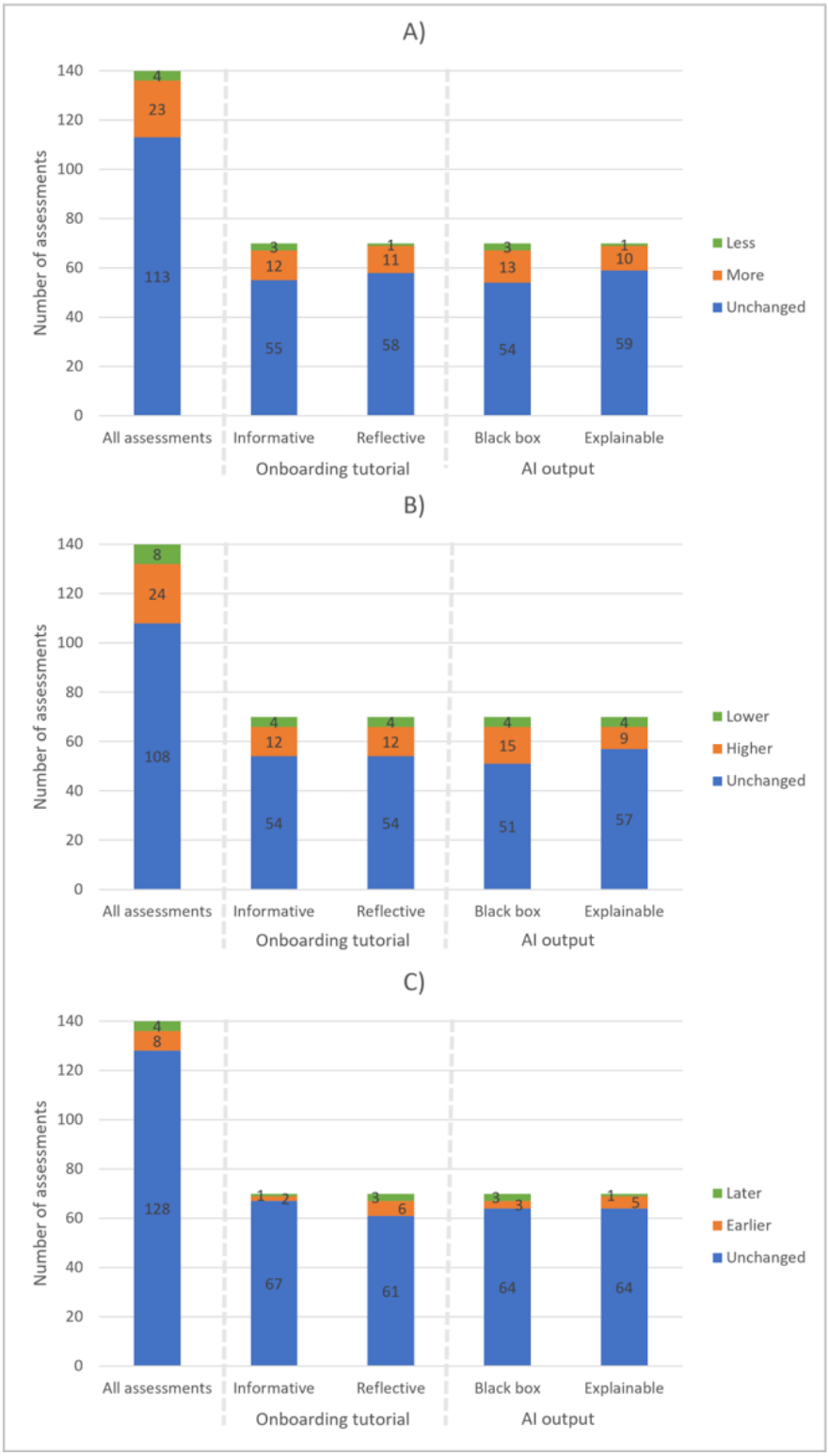
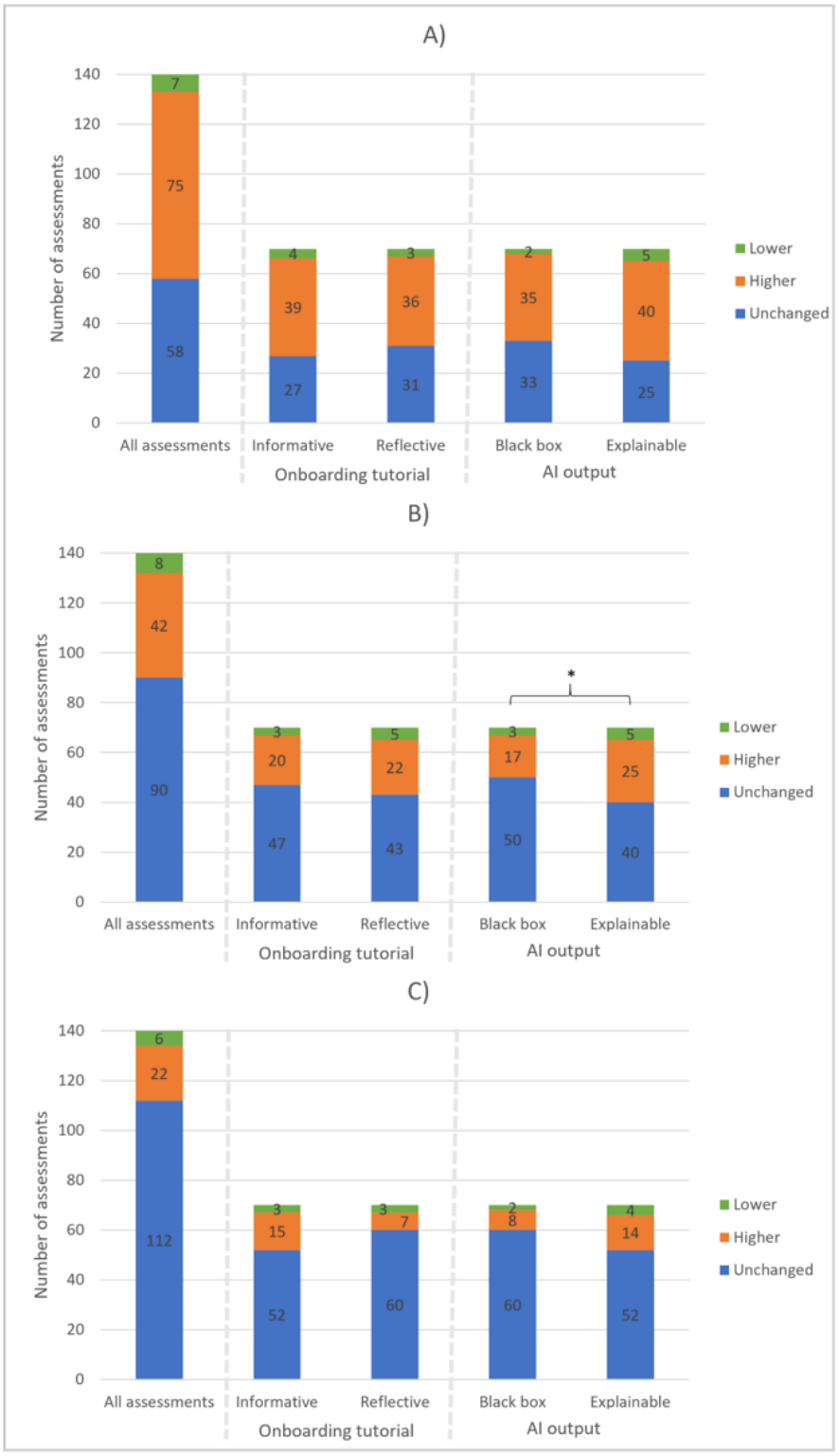


Figure 7. Bar graph showing the changes in the radiologist' confidence in their assessments; (A) Confidence reported nodules, (B) Confidence malignancy probability, (C) Confidence follow-up advice after viewing the recommendations from the artificial intelligence–based computer-aided detection or diagnosis 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: β=–0.245; *P*=.003 third model: β=–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: β=0.369; *P*<.001; third model: β=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: β=0.277; *P*=.02; third model: β=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)

[a]CT: computed tomography.

[b]AI: 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

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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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]
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] [Medline: 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] [Medline: 33706112]
12. Ewals LJS, van der Wulp K, van den Borne BEEM, Pluyter JR, Jacobs I, Mavroeidis 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] [Medline: 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]
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]
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]
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] [Medline: 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, Schmorrow 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]
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]
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] [Medline: 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] [Medline: 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]
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]

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] [Medline: 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]
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]
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=b8eda9593fbcb63b7ced1866853d9622737533a2 [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]
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]
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]
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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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/NEJMoa1102873] [Medline: 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] [Medline: 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]
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] [Medline: 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]
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]
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] [Medline: 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] [Medline: 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] [Medline: 33937824]
47. van Hartskamp 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] [Medline: 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] [Medline: 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] [Medline: 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
PMID:38875574

Original Paper

Perceptions of Family Physicians About Applying AI in Primary Health Care: Case Study From a Premier Health Care Organization

Muhammad Atif Waheed[1], MBBS, MRCS, MRCGP, DPD, MBA; Lu Liu[2], PhD

[1]Primary Health Care Corporation, Doha, Qatar

[2]Bath 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

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 helps 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 Laï 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 α=0.873); (3) family physician perspective on administrative management of AI (10 items, Cronbach α=0.916); (4) family physician's perspective on public health management of AI (9 items, Cronbach α=0.930); and (5) family physicians' perspectives on AI challenges, ethical ramifications, and impact on job roles (13 items, Cronbach α=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	*P* 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.

[a]AI: artificial intelligence.

[b]PHCC: Primary Health Care Corporation.

[c]Correlation coefficient of health care tasks and health care service delivery was Spearman rho=0.679 (moderate to strong correlation).

[d]χ^2_2=71.1; N=102.

[e]χ^2_2=48.8; N=102.

[f]χ^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

1. Laï 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] [Medline: 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]
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]
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] [Medline: 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] [Medline: 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] [Medline: 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]
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] [Medline: 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]
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]
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]
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] [Medline: 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]
14. Hamet P, Tremblay J. Artificial intelligence in medicine. Metabolism 2017 Apr;69S:S36-S40. [doi: 10.1016/j.metabol.2017.01.011] [Medline: 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] [Medline: 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]
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] [Medline: 29507784]
18. Liyanage H, Liaw ST, Jonnagaddala J, Schreiber R, Kuziemsky 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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]
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] [Medline: 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]
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] [Medline: 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] [Medline: 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]
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] [Medline: 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] [Medline: 29992371]
39. Meskó B, Hetényi G, Győrffy 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] [Medline: 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] [Medline: 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] [Medline: 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]
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] [Medline: 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]
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]
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]
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] [Medline: 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
PMID:38875531

Original Paper

Evaluating the Diagnostic Performance of Symptom Checkers: Clinical Vignette Study

Mohammad Hammoud[1], PhD; Shahd Douglas[1], MSc; Mohamad Darmach[1], MD; Sara Alawneh[1], MD; Swapnendu Sanyal[1], MSc; Youssef Kanbour[1], 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

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.

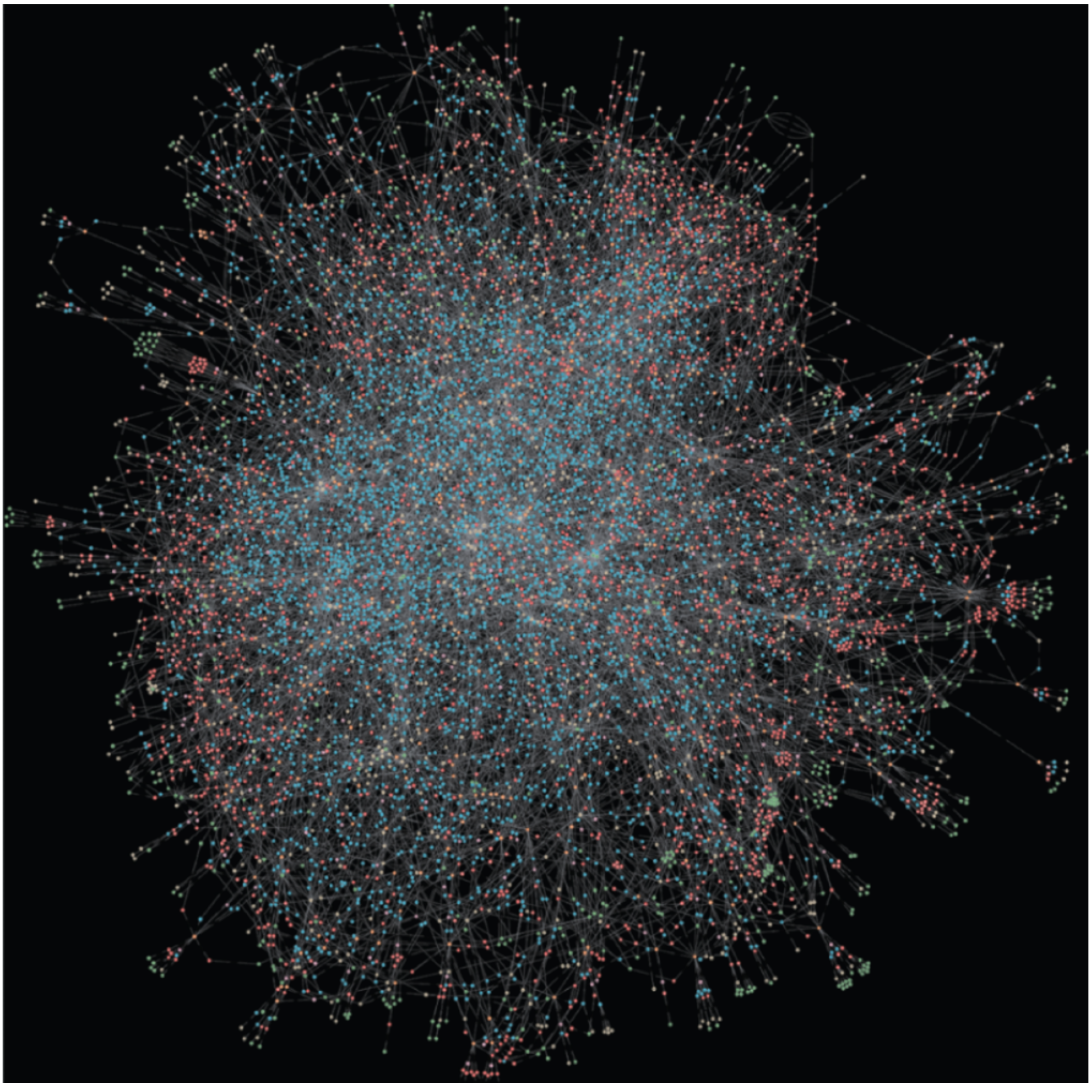
Avey claims to use advanced AI technology [40]. In particular, it involves a diagnostic engine that operationalizes a probabilistic graphical model, namely a Bayesian network. Figure 1 demonstrates the model in action, which was built bottom-up over 4 years specifically for medical diagnosis. In addition, 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

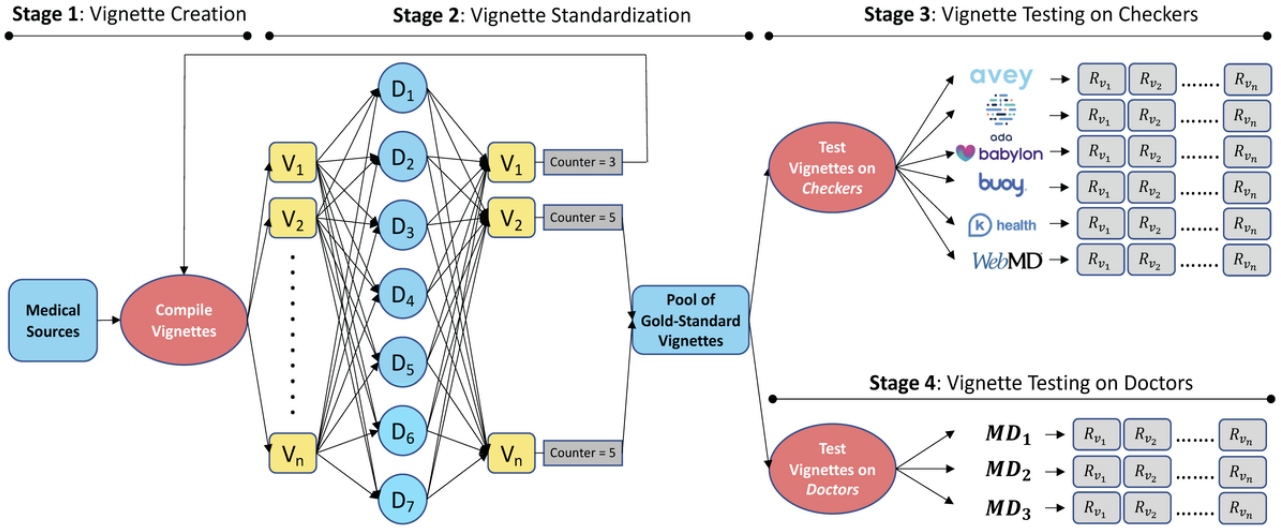
Stages

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≤i≤n; D_j=doctor j, assuming 7 doctors and 1≤j≤7; MD_k=medical doctor k, assuming 3 doctors and 1≤k≤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)

[a]p: number of diseases covered in the body system.

[b]P: total number of diseases covered by the N vignettes.

[c]n: number of vignettes for the corresponding body system.

[d]N: total number of vignettes in our suite.

[e]m: count of vignettes covering common diseases of the corresponding body system.

[f]k: count of vignettes covering less-common diseases of the corresponding body system.

[g]ENT: ear, nose, and throat.

[h]GI: 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 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, *M*1), among the first 3 diseases (ie, *M*3), or among the first 5 diseases (ie, *M*5) of their differential list. For each symptom checker and doctor, we report the percentages of vignettes that fulfill *M*1, *M*3, and *M*5. The mathematical definitions of *M*1, *M*3, and *M*5 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	 , 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	 , 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	 , 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	 , where N is the number of vignettes and  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	 , where N is the number of vignettes and  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	 , where *average recall* and *average precision* 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	 , assuming N vignettes, n number of diseases in a gold standard vignette v, and $relevance_i$ for the disease at position in v's differential list  , which is computed over the differential list of a doctor or a symptom checker for v. *Gold* DCG_v is defined exactly as DCG_v, but is computed over the gold standard differential list of v

[a]NDCG: 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  , 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)

[a]DDx: 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 *M*1; 22.4%, 114.5%, 123.8%, 118.2%, and 3392% using *M*3; 18.1%, 79.2%, 116.8%, 125%, and 3114.2% using *M*5; 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 *M*1; 22.4%, 112.4%, 94%, 112.9%, and 3257.6% using *M*3; 18.1%, 77.8%, 88.2%, 119.5%, and 3003.4% using *M*5; 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.

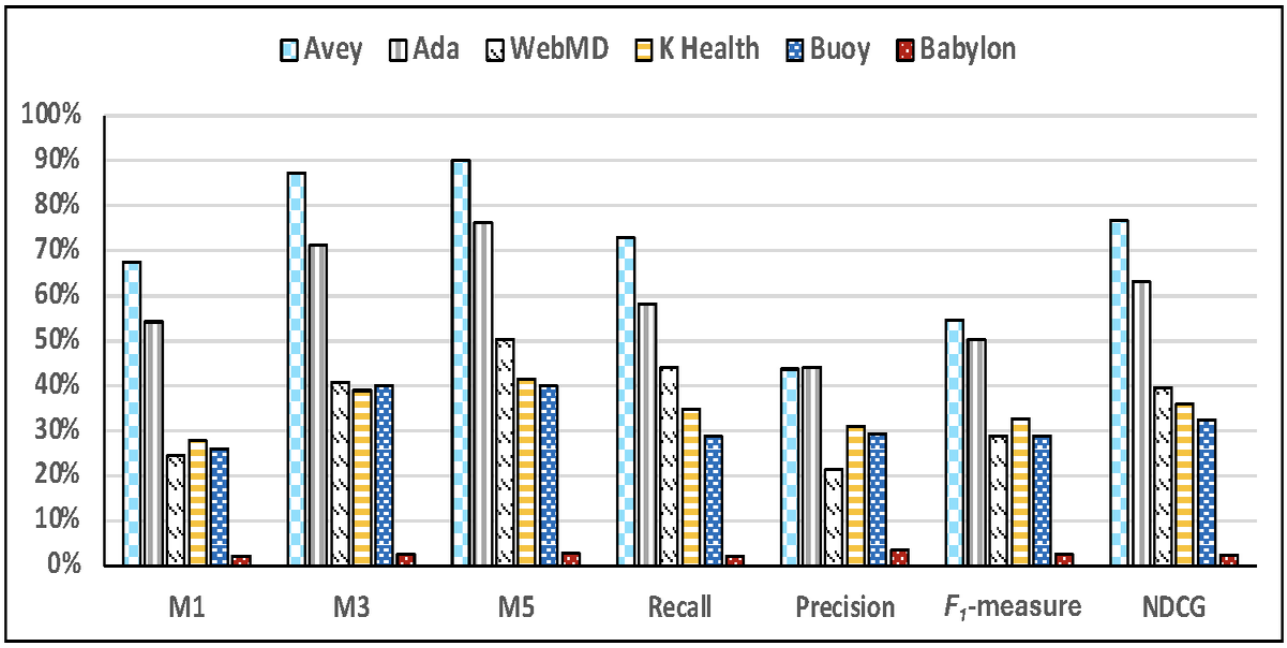
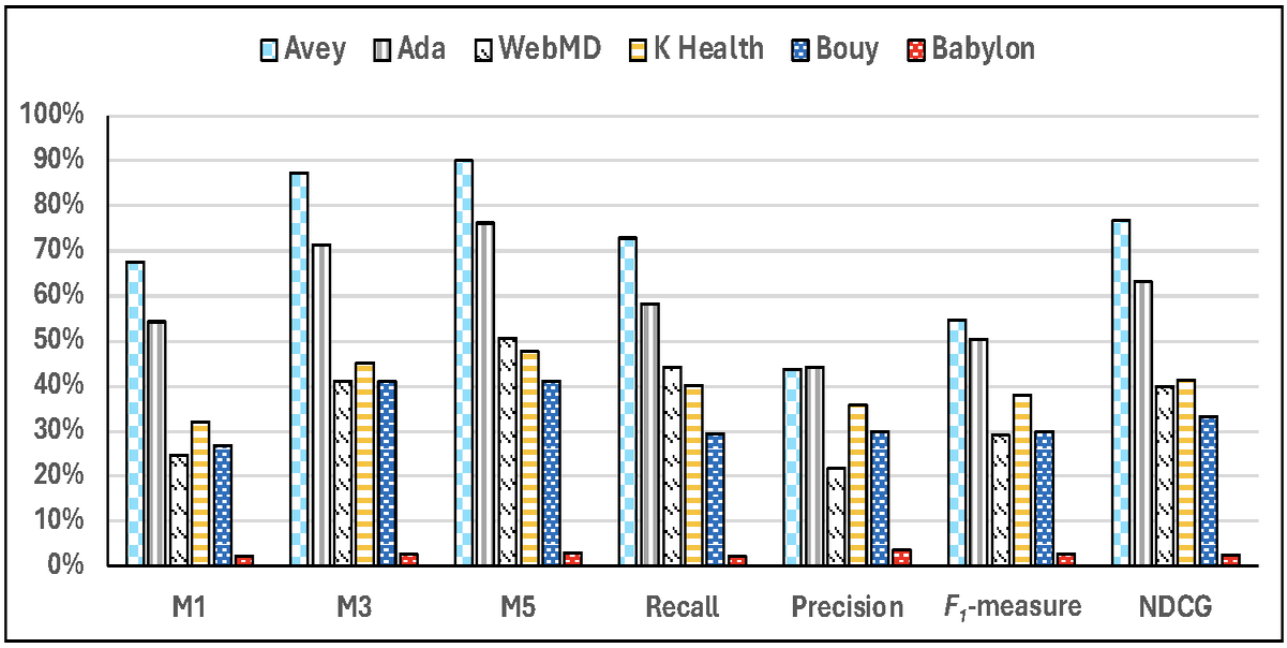


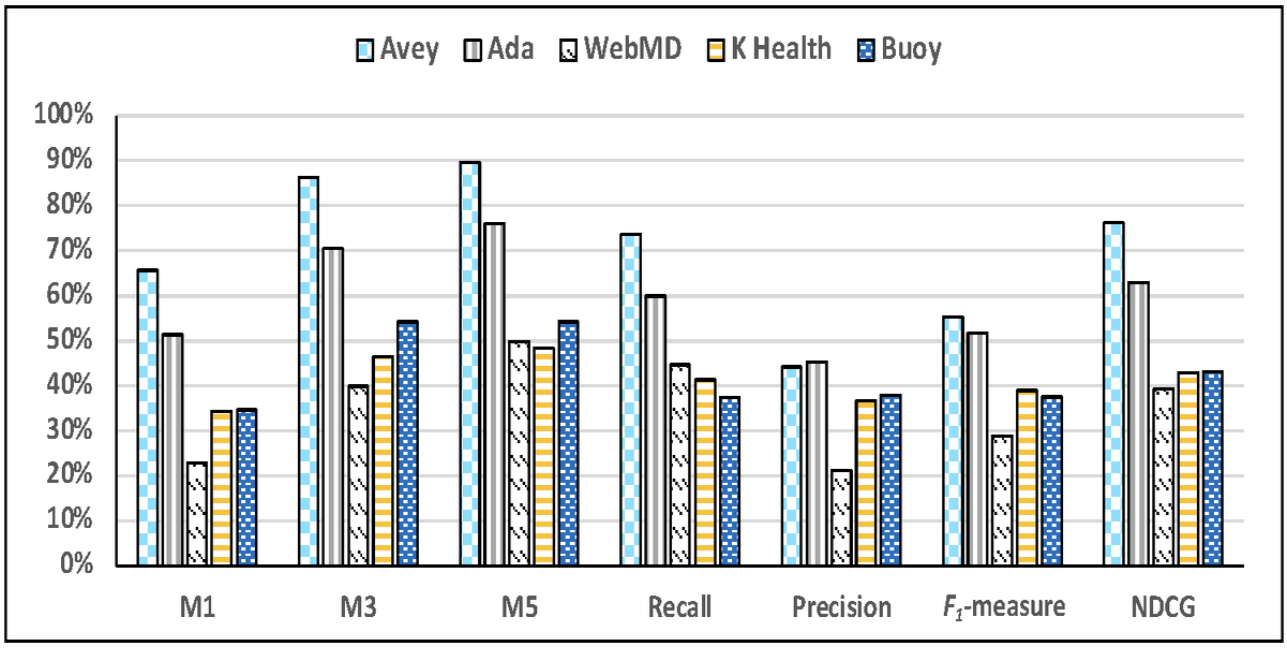
Figure 4. Accuracy results considering for each symptom checker only the succeeded vignettes, with or without differential diagnoses. NDCG: Normalized Discounted Cumulative Gain.



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 *M*1; 22.4%, 116.3%, 85.6%, and 59.2% using *M*3; 18%, 80.1%, 85.7%, and 65.5% using *M*5; 23%, 64.9%, 78.5%, and 97.1% using recall; 7.2%, 92.7%, 42.2%, and 47.1% using F_1-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 *M*1, *M*3, *M*5, recall, precision, F_1-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 *M*1, *M*3, *M*5, recall, precision, F_1-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 F_1-measure but a better performance in terms of *M*1, *M*3, *M*5, NDCG, and recall. More precisely, the doctors outperformed Avey by averages of 37.1% and 1.2% using precision and F_1-measure, whereas Avey outpaced them by averages of 10.2%, 20.4%, 23.4%, 56.4%, and 25.1% using *M*1, *M*3, *M*5, recall, and NDCG, respectively.

Table 4. Accuracy results (%) of 3 medical doctors (MDs), MD_1, MD_2, and MD_3, with an average experience of 16.6 years.

Doctors	M1	M3	M5	Recall	Precision	F_1-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

[a]NDCG: 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 *M*1, 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 *M*1. In contrast, under *M*1, 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 *M*1. 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 MD_1, MD_2, and MD_3) 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 NDCG[a]	Avey, MD_3, Ada, MD_2, MD_1, WebMD, K Health, Buoy, and Babylon	74.2	23	21.3	9

[a]NDCG: 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 *M*1 ranging from 5% to 50% and *M*20 (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 *M*1 and 84.3% vs 51.2% along *M*3). 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 *M*1, *M*3, and *M*5, respectively.

None of the symptom checkers in the study by Gilbert et al [28] outperformed general practitioners but Ada came close, especially in *M*3 and *M*5. 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 *M*1 and from 30% to 81% using *M*10 (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 *M*5.

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], Mediktor [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 Mohmmad 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, Quaye 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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, Smeele 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] [Medline: 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] [Medline: 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]
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] [Medline: 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] [Medline: 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] [Medline: 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]
25. Kasteleyn MJ, Versluis A, van Peet P, Kirk UB, van Dalfsen 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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]
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]
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]
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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 22032458]
62. Hennemann S, Kuhn S, Witthöft 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] [Medline: 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] [Medline: 24434479]
64. Munsch N, Martin A, Gruarin S, Nateqi J, Abdarahmane 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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 MI. Variational probabilistic inference and the QMR-DT network. J Artif Intell Res 1999 May 01;10(1999):291-322. [doi: 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]
77. Miller RA, Pople HEJ, Myers JD. INTERNIST-I, an experimental computer-based diagnostic consultant for general internal medicine. In: Reggia JA, Tuhrim 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]
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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 37266983]
90. Gottliebsen 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] [Medline: 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] [Medline: 34444182]
92. Torjesen I. Patients find GP online services "cumbersome," survey finds. BMJ 2019 Jul 22;366:l4800. [doi: 10.1136/bmj.l4800] [Medline: 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
PMID:38875676

Viewpoint

Toward Clinical Generative AI: Conceptual Framework

Nicola Luigi Bragazzi[1], MPH, MD, PhD; Sergio Garbarino[2], MD, PhD

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

[2]Department 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

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.

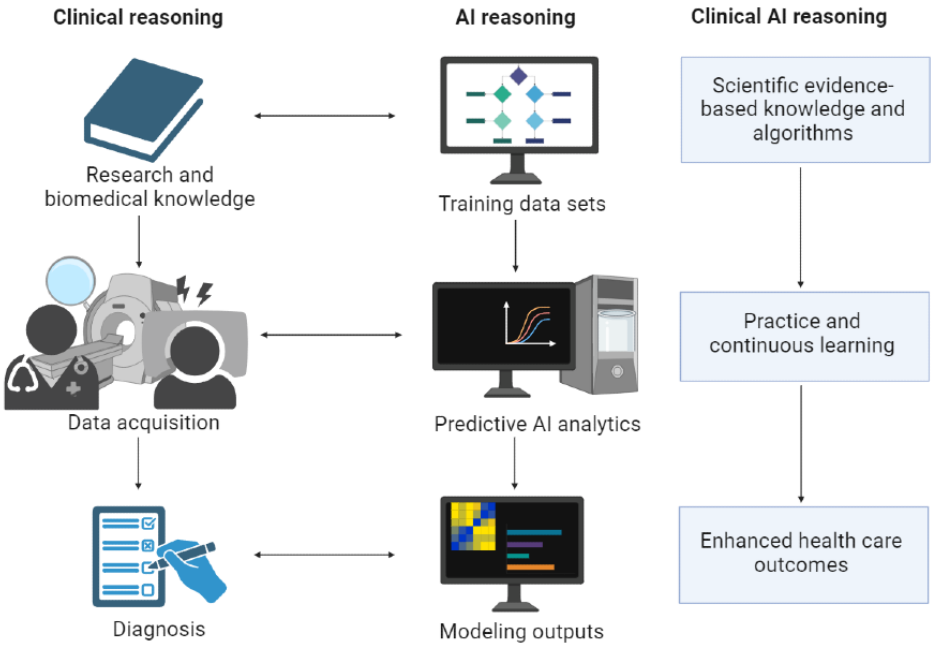
Addressing these challenges requires a multifaceted approach, including improving data set quality and diversity, refining model architectures, and incorporating mechanisms for fact checking and validation. Moreover, developing methodologies for the model to express uncertainty or request clarification 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 parallelisms, 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	• Comprehensive evaluation • Real-world relevance • Assessment of contextual understanding and probabilistic reasoning	• Complex to design • Resource intensive • Potential bias in test creation
Historical data comparison	• Real-world applicability • Evidence-based evaluation • Objective benchmarking	• Dependent on data quality • Historical bias • May not capture AI's[a] potential for novel insights
Expert consensus	• Leverages human expertise • Valuable in complex cases • Incorporates ethical judgment	• Subjective • Time-consuming • Potential for expert bias
Cross-discipline validation	• Comprehensive evaluation from multiple perspectives • Mitigates the risk of siloed decision-making	• Coordination challenges • Requires broad expert availability
Rare or complex simulation and scenario testing	• Reveals AI's capabilities in handling diversity • Can identify areas for innovation	• Potentially limited by available data • Resource intensive
False myth	• Tests AI's current knowledge base • Assesses ability to discern evidence-based information	• Requires careful selection of myths • Risk of reinforcing incorrect information
Challenging (or controversial) question	• Evaluates AI's handling of ambiguity and complexity • Assesses balance of different viewpoints	• Subjective evaluation criteria • Depends on quality of input questions
Real-time monitoring	• Direct insight into practical impact • Simulates real-world testing	• Requires controlled clinical environment • Ethical concerns with experimental use
Algorithm transparency and audit	• Enhances trust and understanding • Facilitates regulatory compliance	• Complexity for end users • Risk of exposing proprietary information
Feedback loop	• Ensures continuous improvement • Adapts to changing medical knowledge	• Requires ongoing effort and resources • Dependence on quality of feedback
Ethical and legal review	• Safeguards patient rights • Ensures adherence to ethical guidelines	• Time-consuming • Needs multidisciplinary expertise

[a]AI: 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.

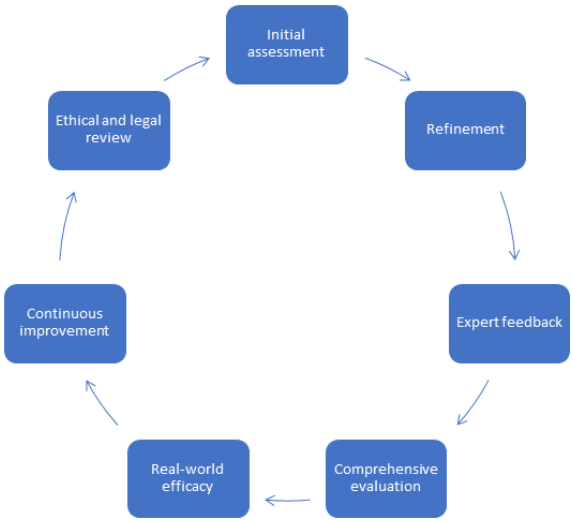
Table 2. Overview of the layered integrative approach for evaluating artificial intelligence (AI) in health care, delineating the structured, multistage framework for the comprehensive assessment and continuous improvement of AI systems.

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] [Medline: 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] [Medline: 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, Berchialla 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]
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] [Medline: 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] [Medline: 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]
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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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]

16. Savcisens 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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]
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] [Medline: 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 Andreoletti, 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
PMID:38875592

Research Letter

What Is the Performance of ChatGPT in Determining the Gender of Individuals Based on Their First and Last Names?

Paul Sebo[1], 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

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)

[a]The total number of physicians does not add to 6131 because of missing values (no assignments for 916 physicians).

[b]If 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≤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" (κ>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)	*P* 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)		

[a]Interrater agreement between ChatGPT-3.5 and ChatGPT-4 (for the first round): Cohen κ=0.9768, 95% CI 0.9715-0.9822, *P*<.001.

[b]Interrater agreement between the first and second rounds for each version.

[c]Performance metrics: errorCoded=0.01223, errorCodedWithoutNA=0.01223, naCoded=0, and errorGenderBias=–0.00636.

[d]Performance metrics: errorCoded=0.01354, errorCodedWithoutNA=0.01354, naCoded=0, and errorGenderBias=–0.00440.

[e]Performance metrics: errorCoded=0.01501, errorCodedWithoutNA=0.01501, naCoded=0, and errorGenderBias=–0.00620.

[f]Performance 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 ≤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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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
PMID:

Original Paper

Leveraging Machine Learning to Develop Digital Engagement Phenotypes of Users in a Digital Diabetes Prevention Program: Evaluation Study

Danissa V Rodriguez[1], PhD; Ji Chen[1], PhD; Ratnalekha V N Viswanadham[1], PhD; Katharine Lawrence[1,2], MPH, MD; Devin Mann[1,2], MS, MD

[1]New York University Grosman School of Medicine, New York, NY, United States

[2]New York University Langone Health, New York, NY, United States

Corresponding Author:
Danissa V Rodriguez, PhD
New York University Grosman 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

(JMIR AI 2024;3:e47122) doi:10.2196/47122

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].

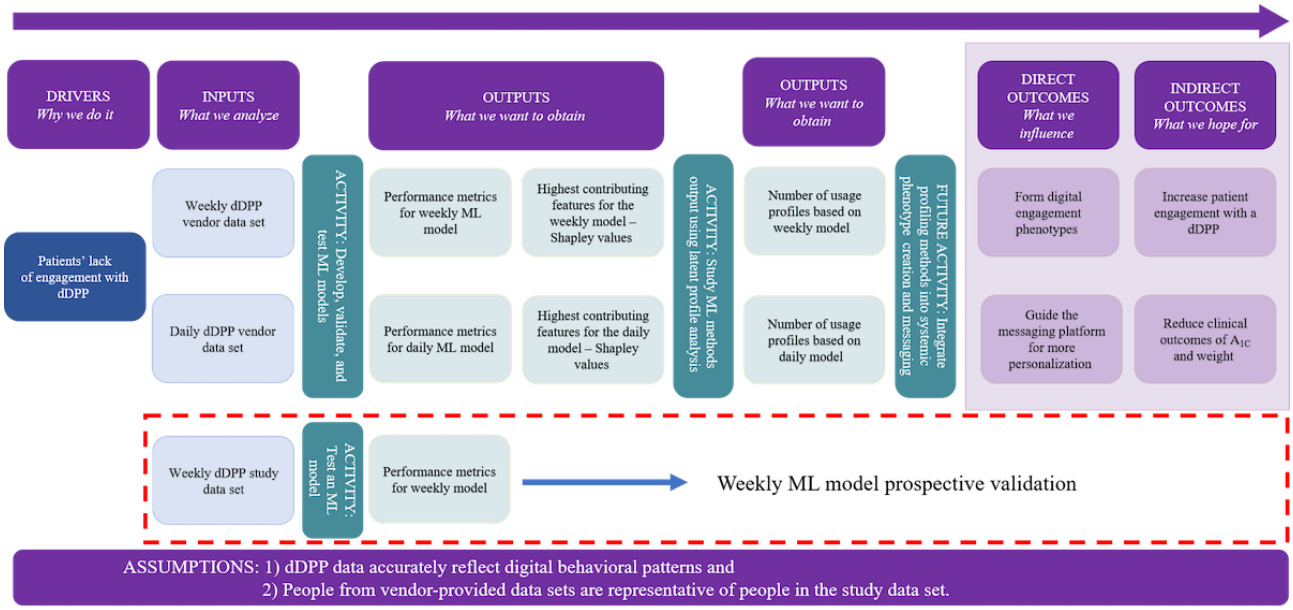
Prior work from this team involved developing a personalized automatic messaging system (PAMS) that leveraged an evidence-based engagement algorithm to deliver tailored behavior change theory–supported SMS text messaging to support users engaging with a commercial app-based dDPP. The study returned promising 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/m^2 (22 kg/m^2 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)

[a]dDPP: digital diabetes prevention program.

[b]N/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.

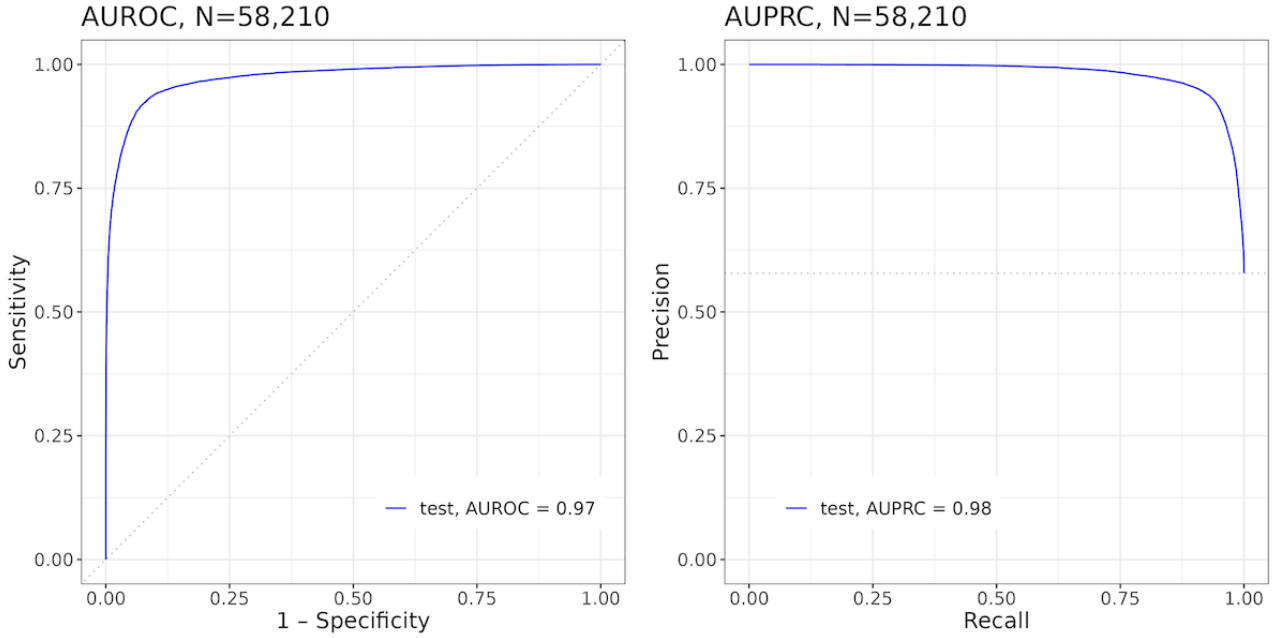
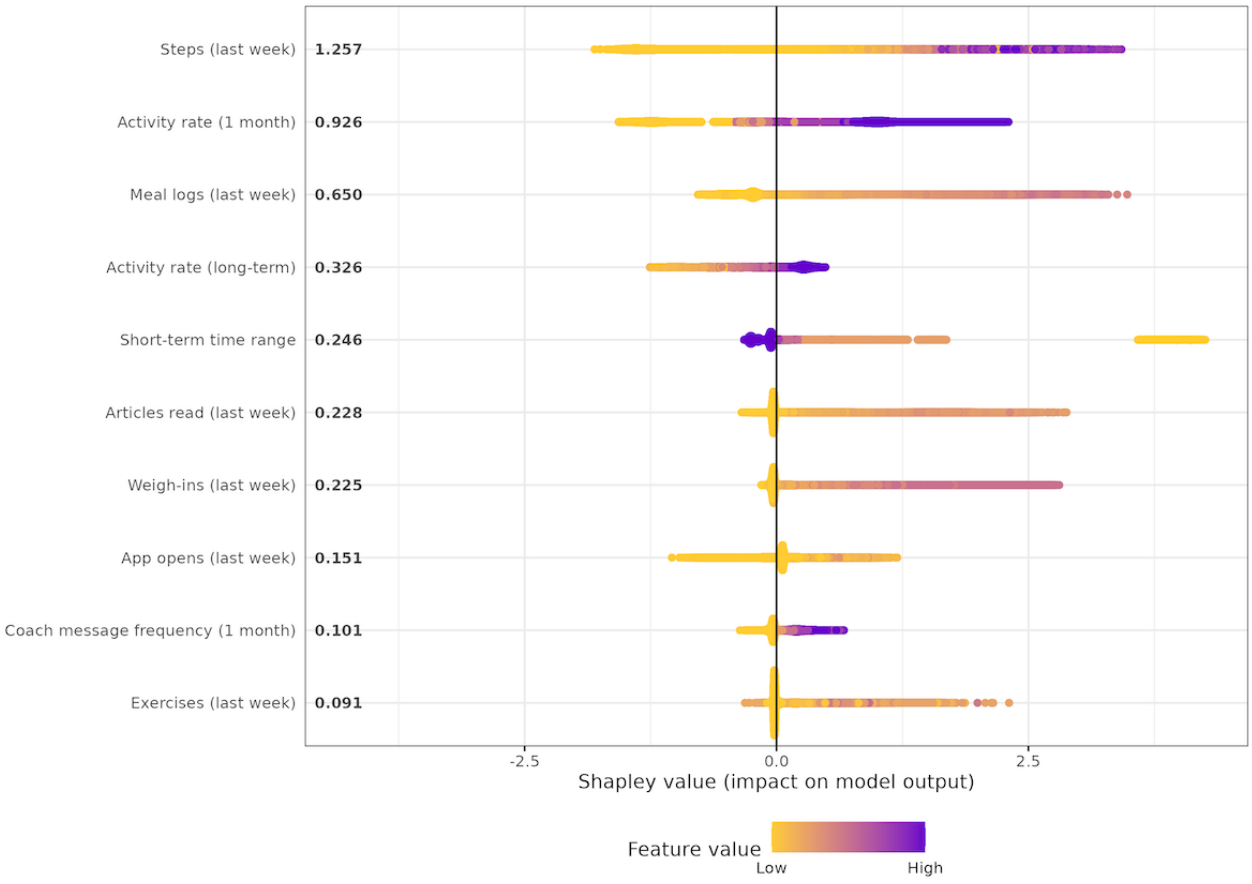


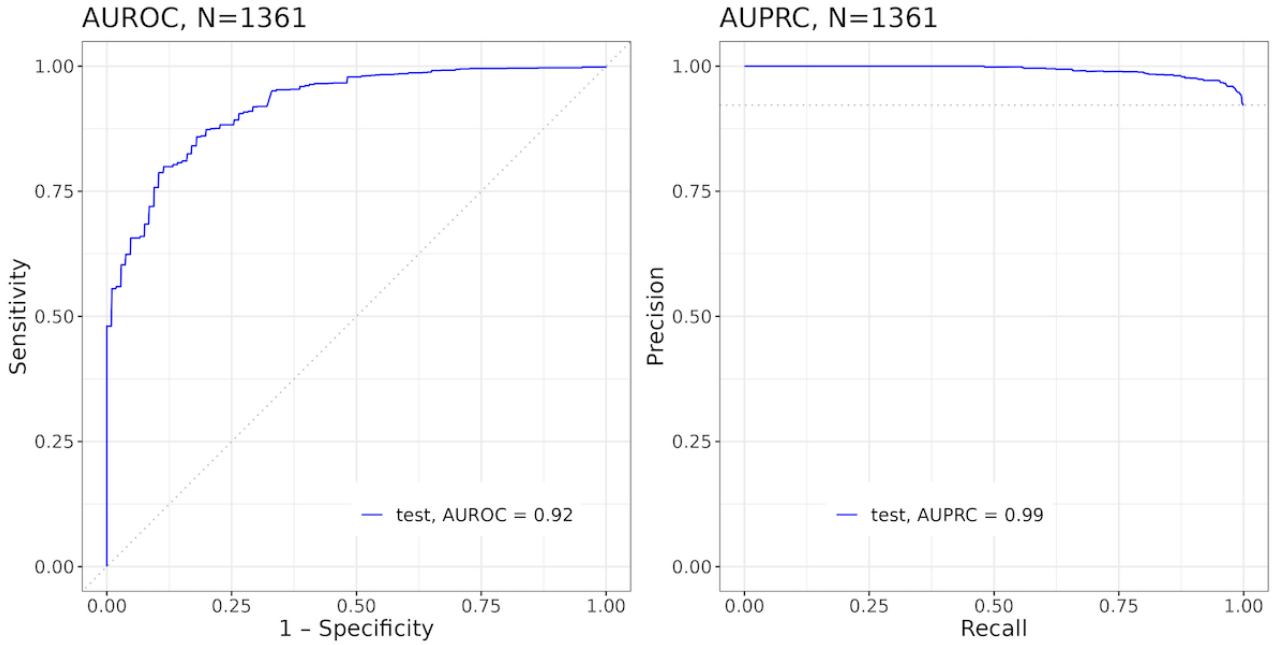
Figure 3 displays the distribution of the 10 covariates with the highest calculated global mean Shapley value (ie, which variables have the strongest predictive power, regardless of negative or positive impact, on the user's engagement with the dDPP). A higher magnitude of the Shapley value (ie, further 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.

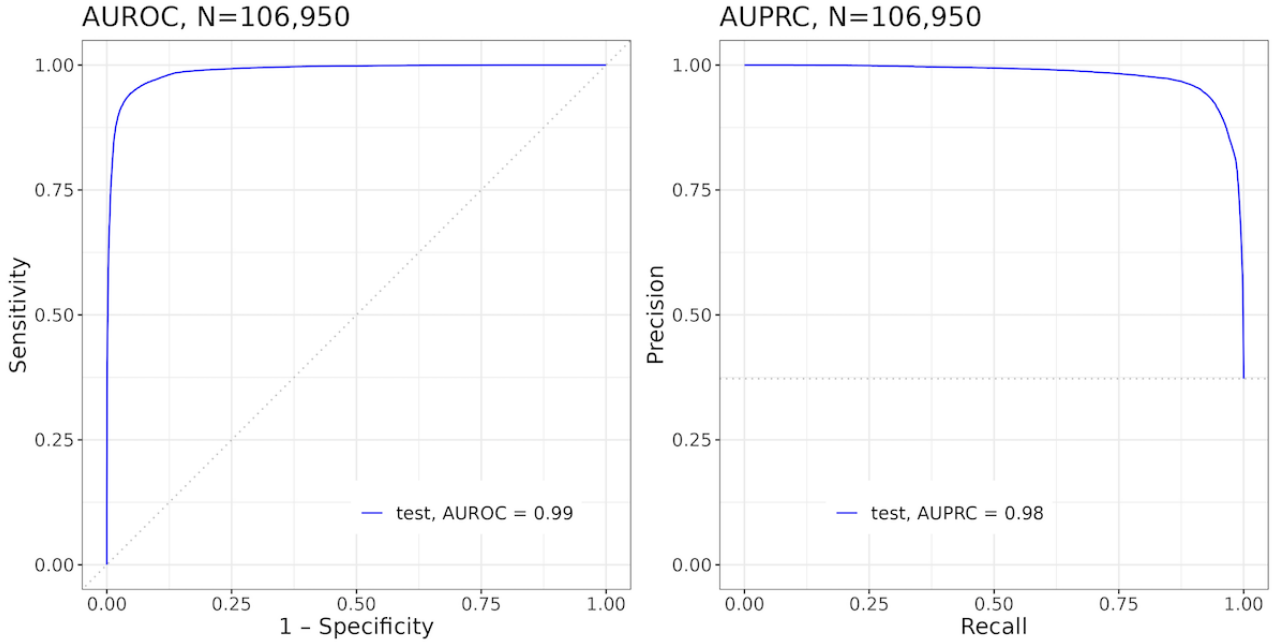
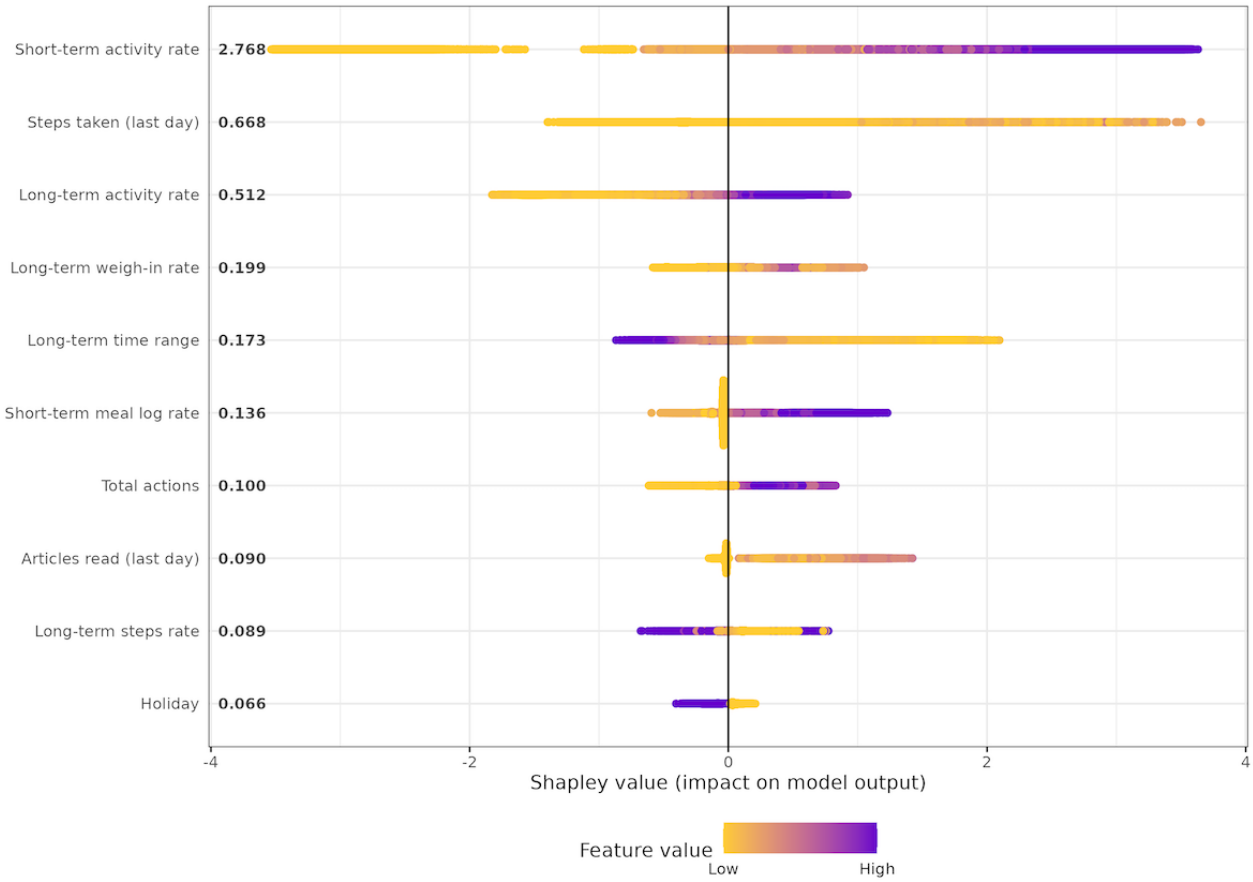


Figure 6. Shapley values of top 10 features in the "any activity" daily model. Each dot on the plot represents an engagement record and is colored according to the value of the corresponding feature from high (purple) to low (yellow). Features are ranked in descending order from top to bottom on the 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)

[a]AUROC: area under the receiver operating characteristic curve.

[b]AUPRC: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 25965228]
9. Alkhaldi 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] [Medline: 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] [Medline: 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] [Medline: 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]
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] [Medline: 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]
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] [Medline: 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] [Medline: 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]
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/wzftr]
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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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
PMID:38875579

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 Young[1], MD; Chin-Wen Chang[2], PhD; Charles W Scales[3], PhD; Saurabh V Menon[4], BTech; Chantal E Holy[5], MS, PhD; Caroline Adrienne Blackie[6], MS, OD, PhD

[1]Department of Ophthalmology, New York University School of Medicine, New York, NY, United States

[2]Data Science, Johnson & Johnson MedTech, Raritan, NJ, United States

[3]Medical and Scientific Operations, Johnson & Johnson Medtech, Vision, Jacksonville, FL, United States

[4]Mu Sigma Business Solutions Private Limited, Bangalore, India

[5]Epidemiology and Real-World Data Sciences, Johnson & Johnson MedTech, New Brunswick, NJ, United States

[6]Medical 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

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)

[a]ICD-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

[a]AMD: age-related macular degeneration.

[b]OSD: ocular surface disease.

[c]NPDR: nonproliferative diabetic retinopathy.

[d]PDR: 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.

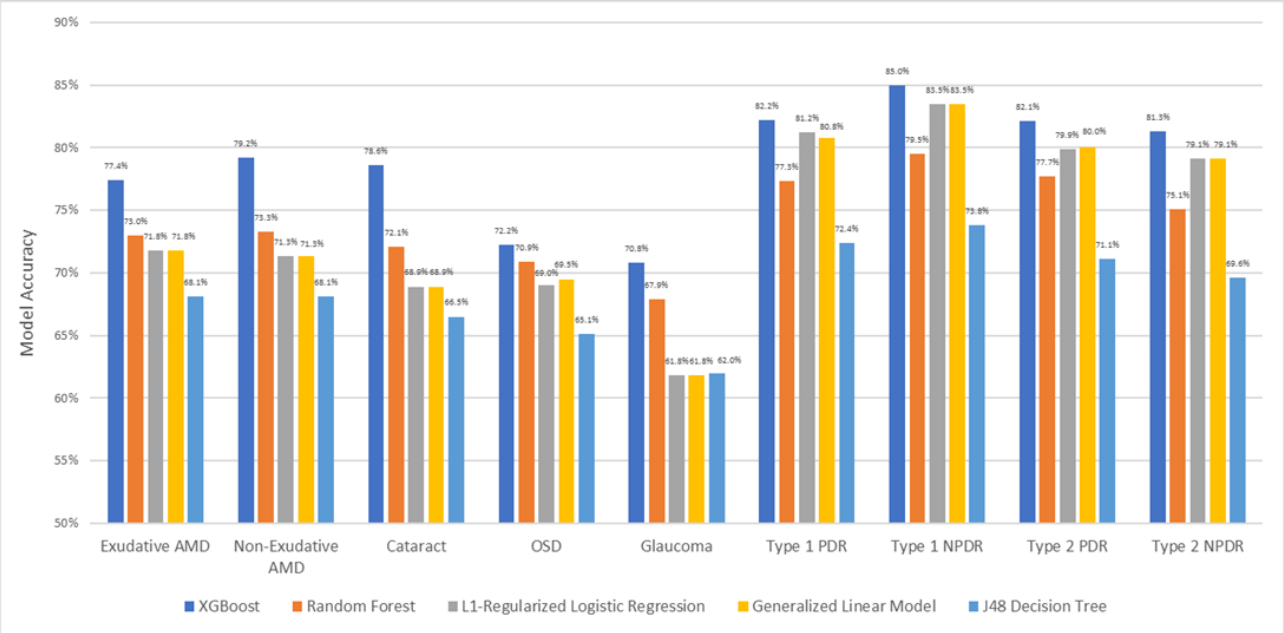


Figure 2. Receiver operating characteristic (ROC) curves illustrating the diagnostic ability of the models for the 9 pathologies. amd: age-related macular degeneration; auc: area under the curve; demgd: dry eye and meibomian gland dysfunction; j48: decision tree; l1: L1-regularized logistic regression; lr: logistic regression without regularization; npdr: nonproliferative diabetic retinopathy; pdr: proliferative diabetic retinopathy; rf: random forest; xgb: Extreme Gradient Boosting.

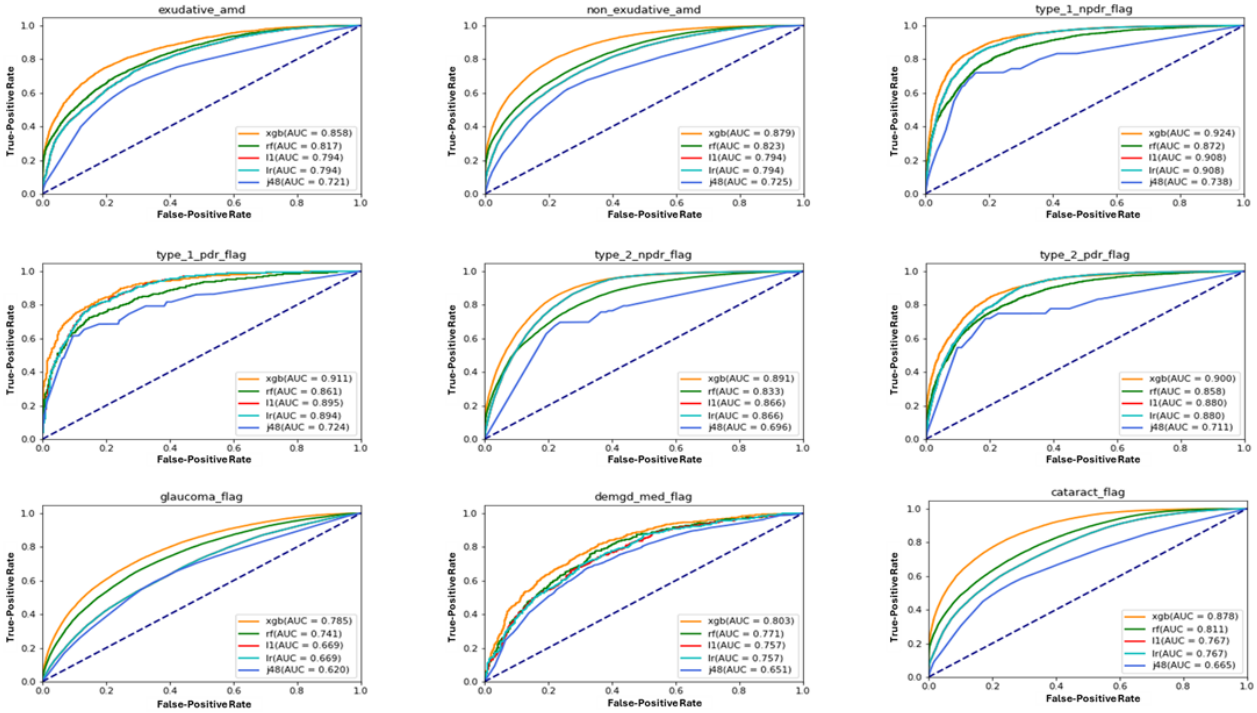


Table 4. Model accuracy, AUC[a], sensitivity, and specificity.

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
OSD[i]				
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

[a]AUC: area under the curve.

[b]XGBoost: Extreme Gradient Boosting.

[c]RF: random forest.

[d]L1-LR: L1-regularized logistic regression.

[e]LR: logistic regression without regularization.

[f]DT: decision tree.

[g]N/A: not applicable.

[h]AMD: age-related macular degeneration.

[i]OSD: ocular surface disease.

[j]NPDR: nonproliferative diabetic retinopathy.

[k]PDR: 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.

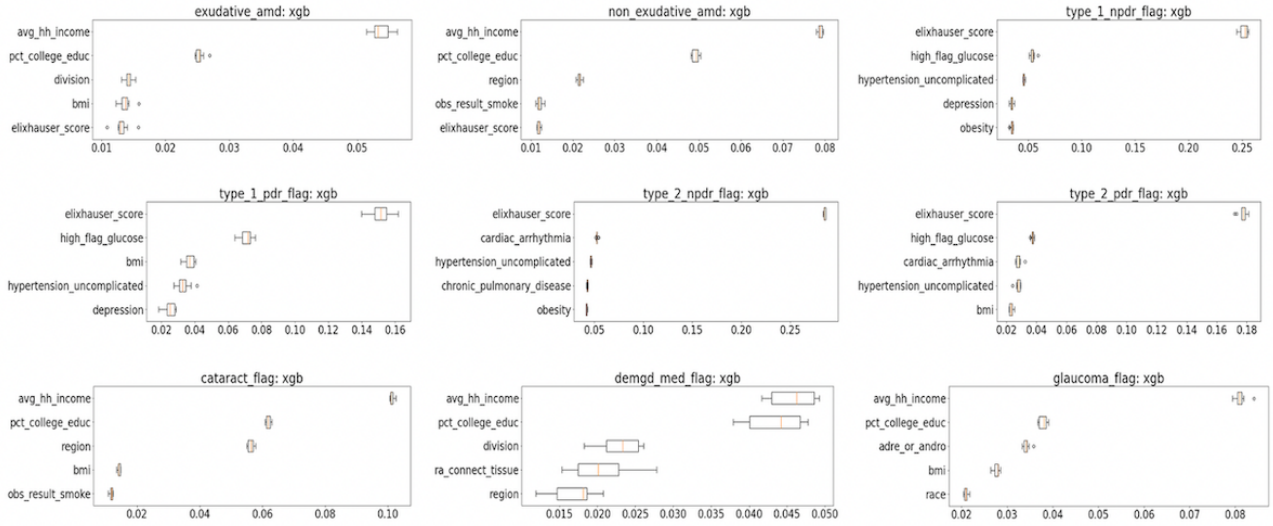


Table 5. Model accuracy and ORs[a] by pathology.

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)

[a]OR: odds ratio.

[b]AMD: age-related macular degeneration.

[c]OSD: ocular surface disease.

[d]PDR: proliferative diabetic retinopathy.

[e]NPDR: 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 disproportionally 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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]
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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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.0000000000005761]
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/archopht.120.6.714] [Medline: 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] [Medline: 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/archopht.1995.01100090057023] [Medline: 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] [Medline: 28855195]
15. Chakravarthy U, Wong TY, Fletcher A, Piault 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] [Medline: 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]
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] [Medline: 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] [Medline: 31481587]
19. Stewart J, Sprivulis 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] [Medline: 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] [Medline: 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]
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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 10861777]
32. Lee S, Lee H, Abbeel P, Ng A. EfficientL1 regularized logistic regression. 2006 Presented at: AAAI'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] [Medline: 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]
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]
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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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/opth.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]
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] [Medline: 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]
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] [Medline: 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] [Medline: 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] [Medline: 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]
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]
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]
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]

51. Khera R, Butte AJ, Berkwits M, Hswen Y, Flanagin 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] [Medline: 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] [Medline: 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
PMID:38875582

Original Paper

A Comparison of Personalized and Generalized Approaches to Emotion Recognition Using Consumer Wearable Devices: Machine Learning Study

Joe Li[1]; Peter Washington[1], 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

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.

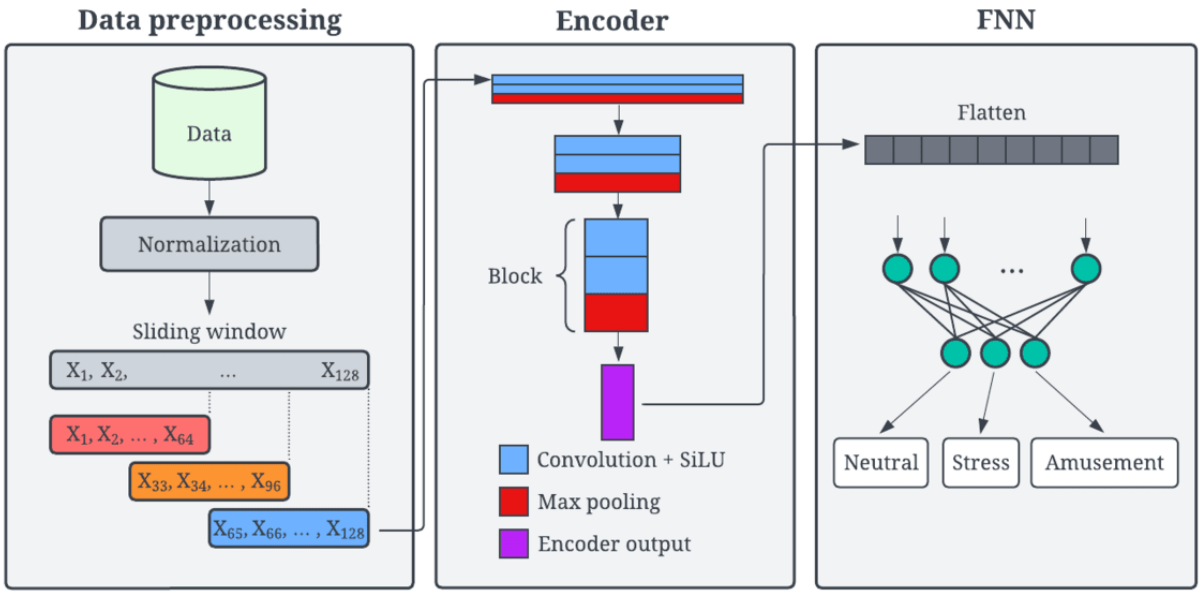
We present 1 personalized and 2 generalized machine learning approaches for the 3-class emotion classification problem (neutral, stress, and amusement) on the Wearable Stress and Affect Detection (WESAD) data set, a publicly available data set that includes both stress and emotion data [18]. The two 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.

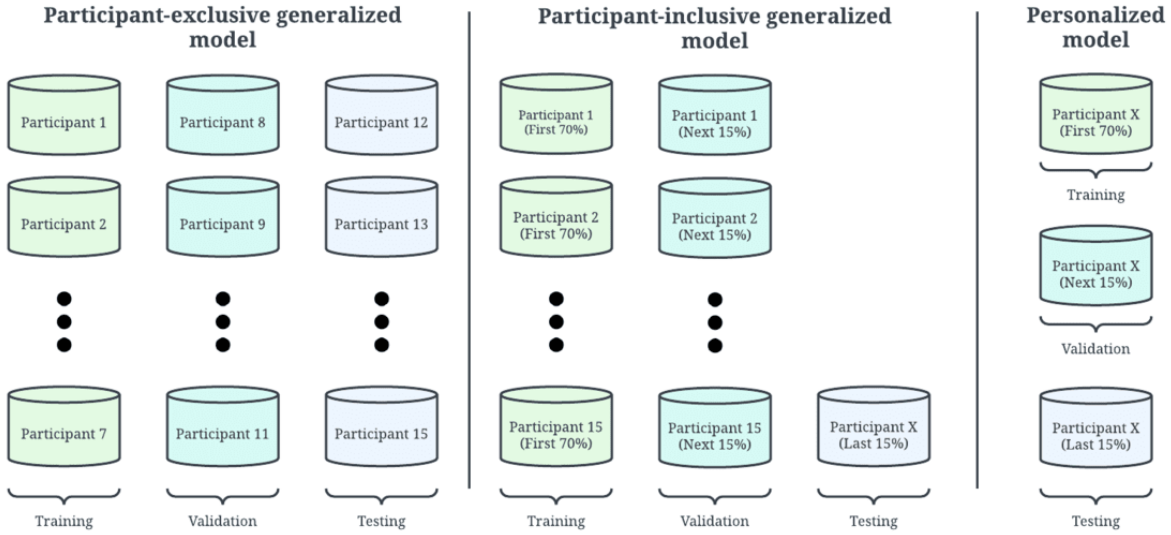
Preprocessing and Partitioning

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]

[a]SiLU: 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	*P* value for accuracy	*P* 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.

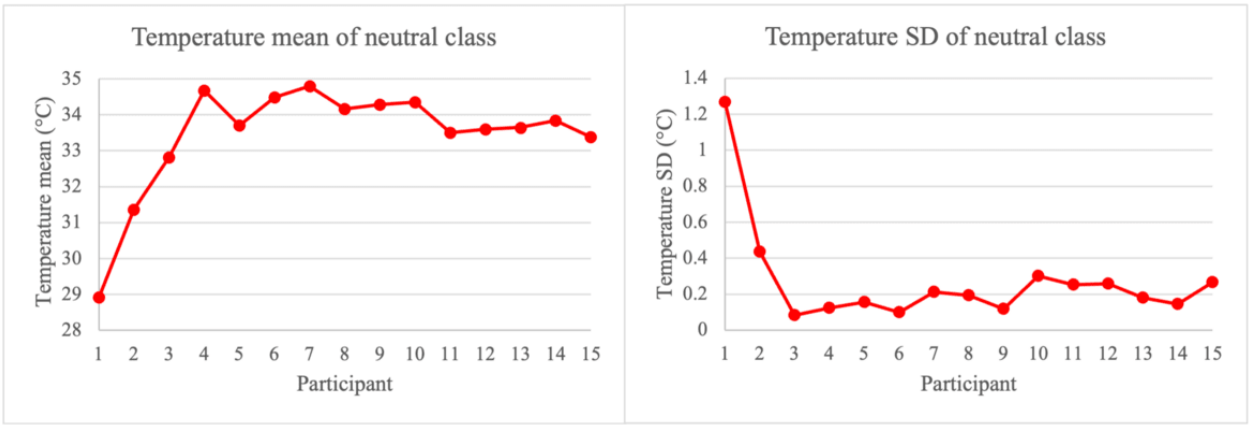


Figure 4. Deviations of mean and SD for subjects 1 and 2 for stress class modalities. EMG: electromyogram.

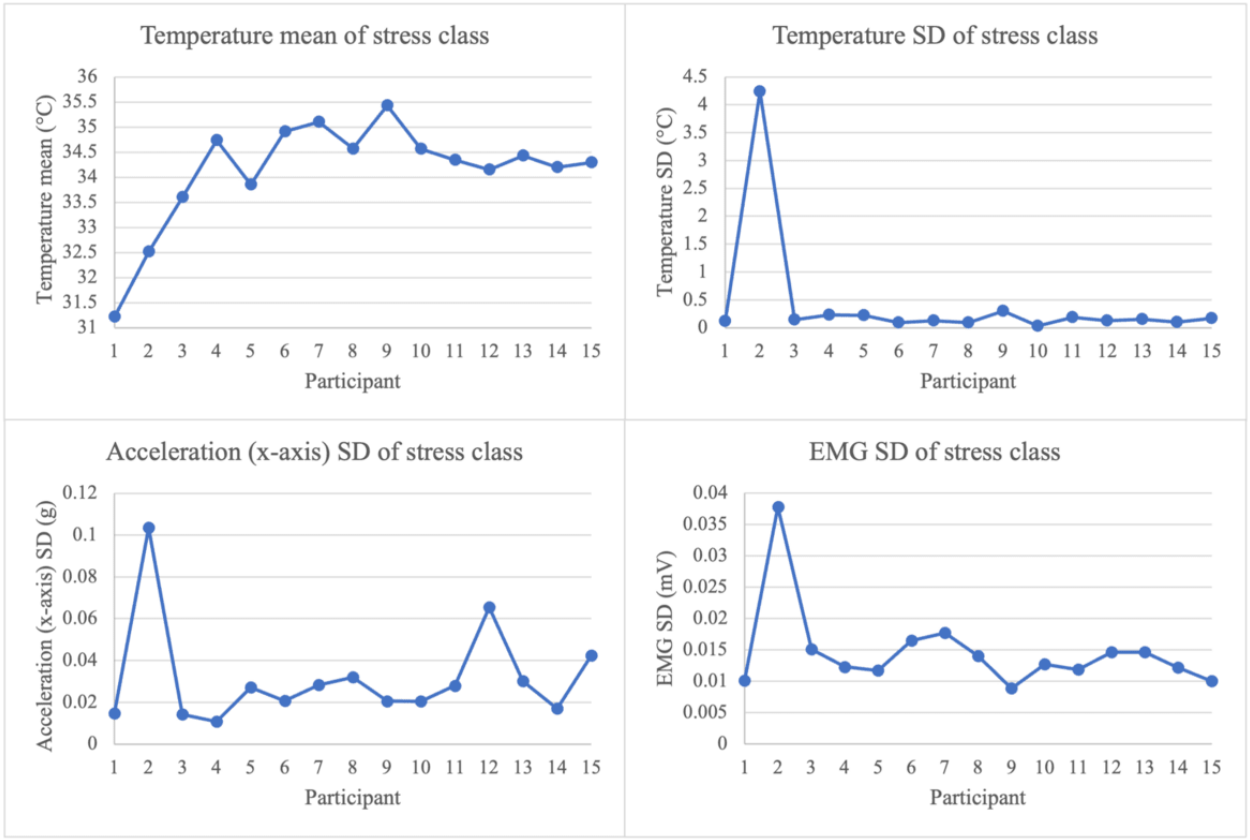


Figure 5. Deviations of mean and SD for subjects 1 and 2 for amusement class modalities.

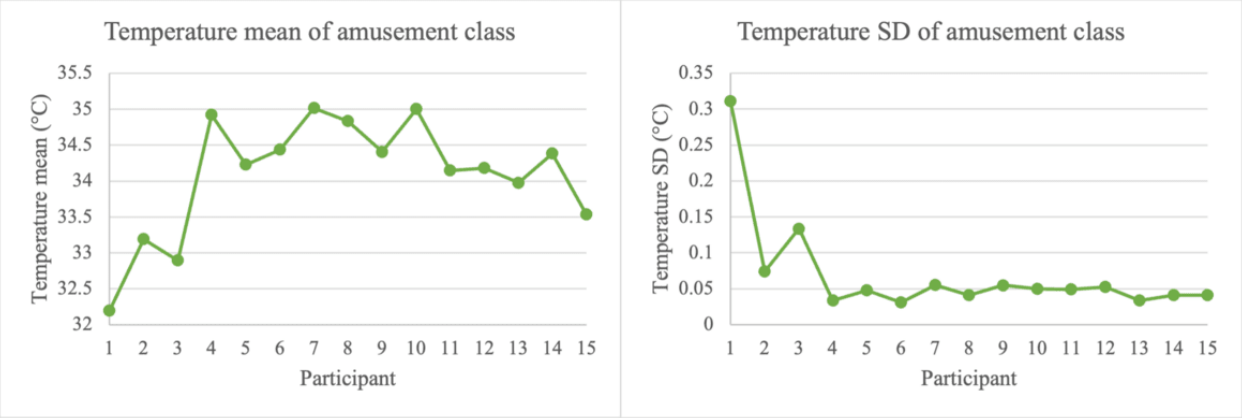


Table 6. Ranges of emotion class distributions per participant.

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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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]
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] [Medline: 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] [Medline: 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] [Medline: 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]

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]
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] [Medline: 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]
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]
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]
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]
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]
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]
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]
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]
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] [Medline: 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]
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]
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] [Medline: 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] [Medline: 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]
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]
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] [Medline: 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]
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] [Medline: 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]
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]
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]
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]
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]
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]
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]
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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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
PMID:38875573

Original Paper

Sample Size Considerations for Fine-Tuning Large Language Models for Named Entity Recognition Tasks: Methodological Study

Zoltan P Majdik[1], PhD; S Scott Graham[2], PhD; Jade C Shiva Edward[2], MA; Sabrina N Rodriguez[3], BS; Martha S Karnes[4], PhD; Jared T Jensen[2], MA; Joshua B Barbour[5], PhD; Justin F Rousseau[6,7], MD, MMSc

[1]Department of Communication, North Dakota State University, Fargo, ND, United States

[2]Department of Rhetoric & Writing, The University of Texas at Austin, Austin, TX, United States

[3]Department of Neurology, The Dell Medical School, The University of Texas at Austin, Austin, TX, United States

[4]Department of Rhetoric & Writing, University of Arkansas Little Rock, Little Rock, AR, United States

[5]Department of Communication, The University of Illinois at Urbana-Champaign, Urbana, IL, United States

[6]Statistical Planning and Analysis Section, Department of Neurology, The University of Texas Southwestern Medical Center, Dallas, TX, United States

[7]Peter 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

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 *chngpt* 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

[a]EPS: entities per sentence.

[b]Individual 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)

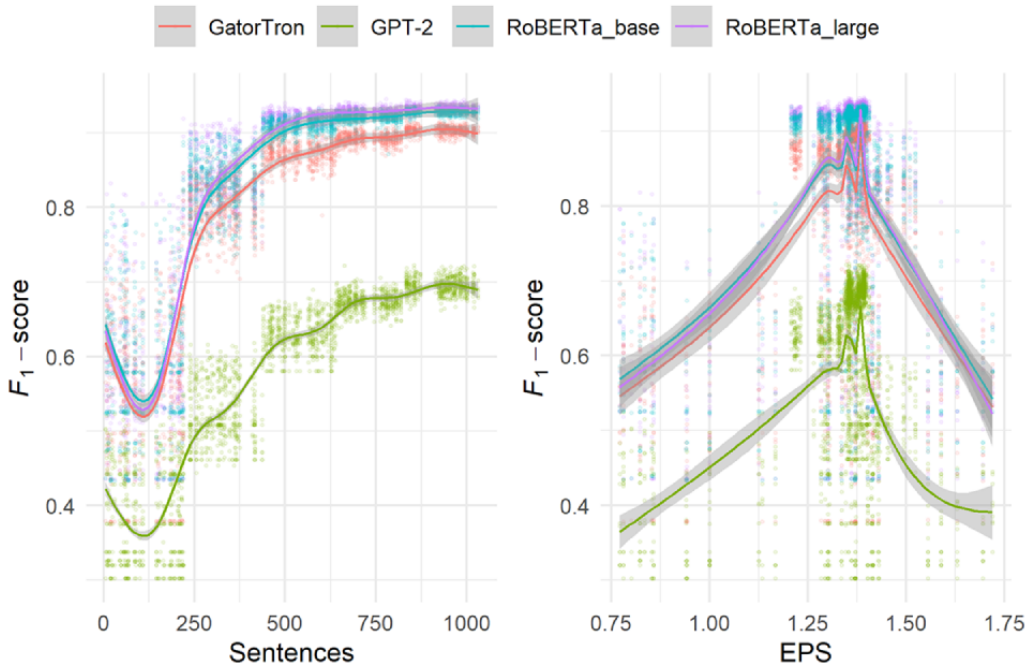
[c]Predictor results are significant at the *P*<.01 level.

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)

[a]EPS: 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 levering 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] [Medline: 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]
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] [Medline: 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]
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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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]
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] [Medline: 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]

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] [Medline: 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] [Medline: 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]
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] [Medline: 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]
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] [Medline: 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]
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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 35673045]
38. Grundy Q, Dunn AG, Bero L. Improving researchers' conflict of interest declarations. BMJ 2020;368:m422. [doi: 10.1136/bmj.m422] [Medline: 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] [Medline: 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]
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]

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] [Medline: 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]
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]
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]

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
PMID:38875593

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 Okubo[2], MD; Ryota Koshu[1], MD; Yuki Uchiyama[2], MD; Akihiro Nomura[3], MD, PhD; Makoto Ito[1], MD, PhD; Yutaka Takumi[2], MD, PhD

[1]Department of Otolaryngology, Head and Neck Surgery, Jichi Medical University, Shimotsuke, Japan

[2]Department of Otolaryngology - Head and Neck Surgery, Shinshu University, Matsumoto, Japan

[3]College of Transdisciplinary Sciences for Innovation, Kanazawa University, Kanazawa, Japan

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

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.

(JMIR AI 2024;3:e58342) doi:10.2196/58342

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.

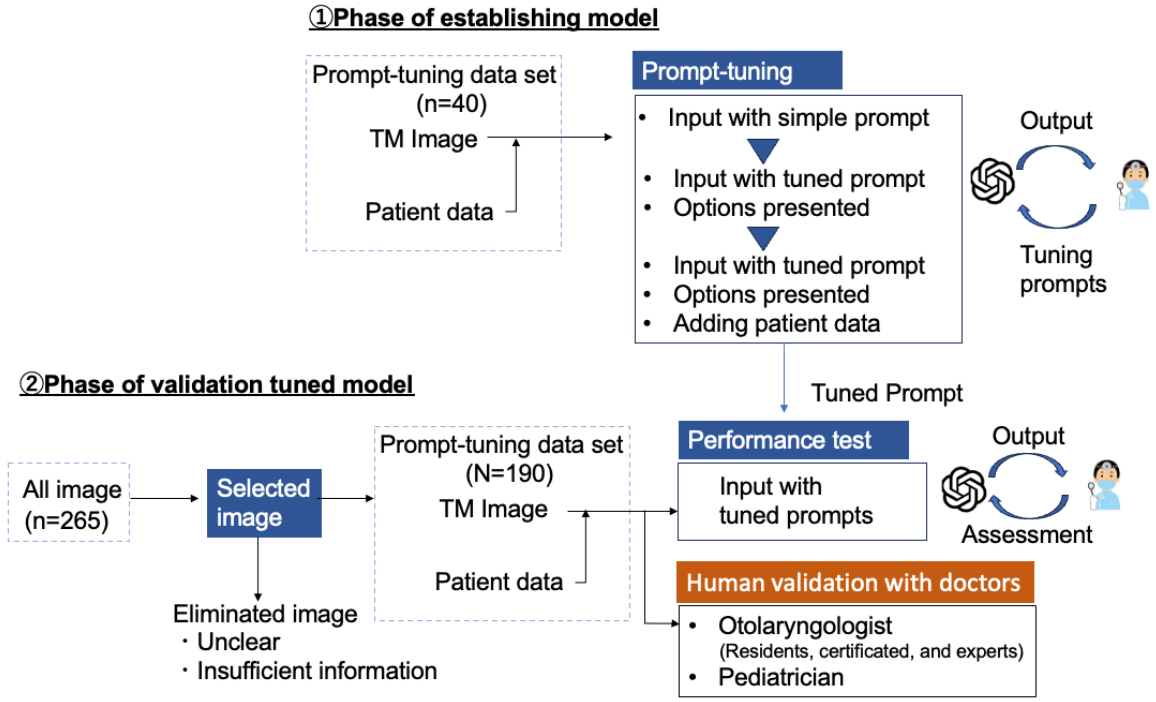
In this study, we aimed to investigate the effectiveness of a multimodal approach using GPT-4V to diagnose middle ear disease. This approach was designed to integrate patient-specific data (age, sex, and chief complaint) with tympanic membrane images to assess the accuracy of the versatile GPT-4V. The 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 Otoscopic 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.

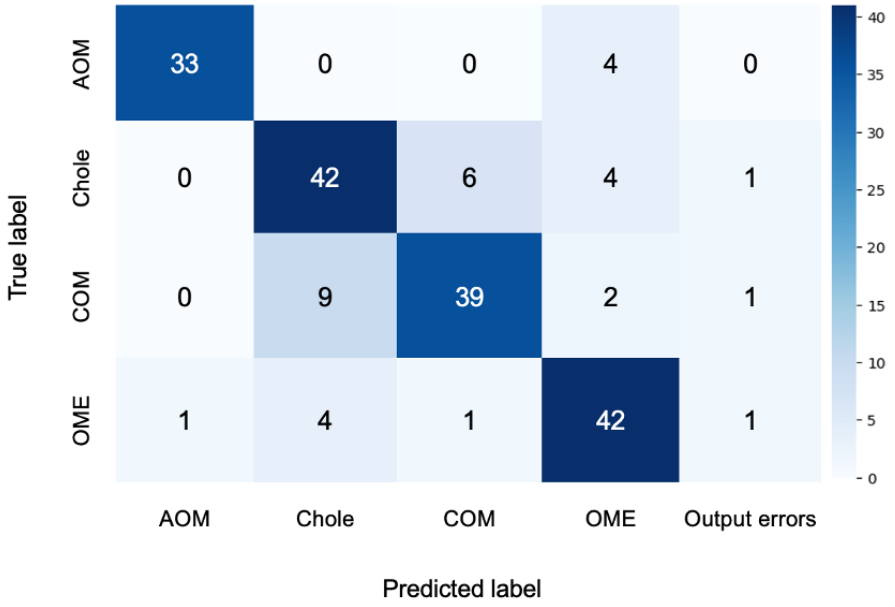
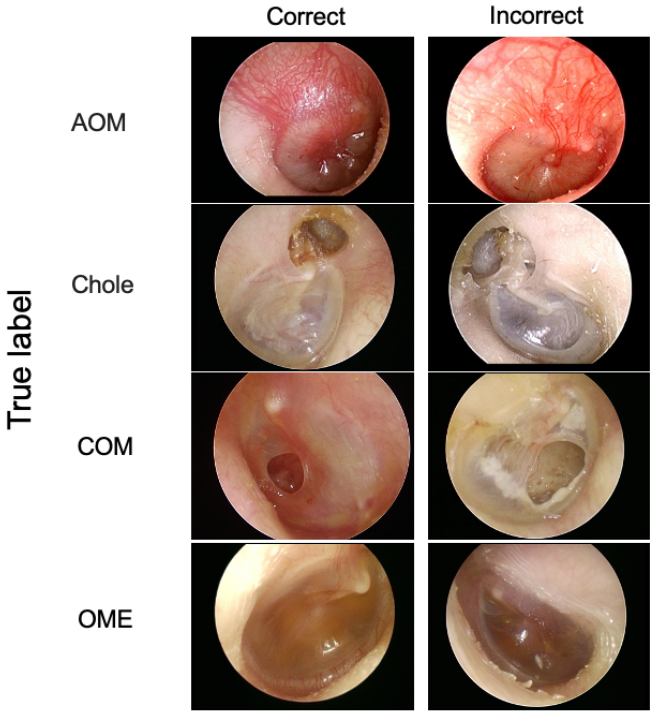


Figure 3. Representative images of correct and incorrect GPT-4V classifications for 4 middle ear diseases. The left side shows the correct images for GPT-4V classification, and the right side shows the incorrect images for GPT-4V. AOM: acute otitis media; chole: middle ear cholesteatoma; COM: chronic 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.

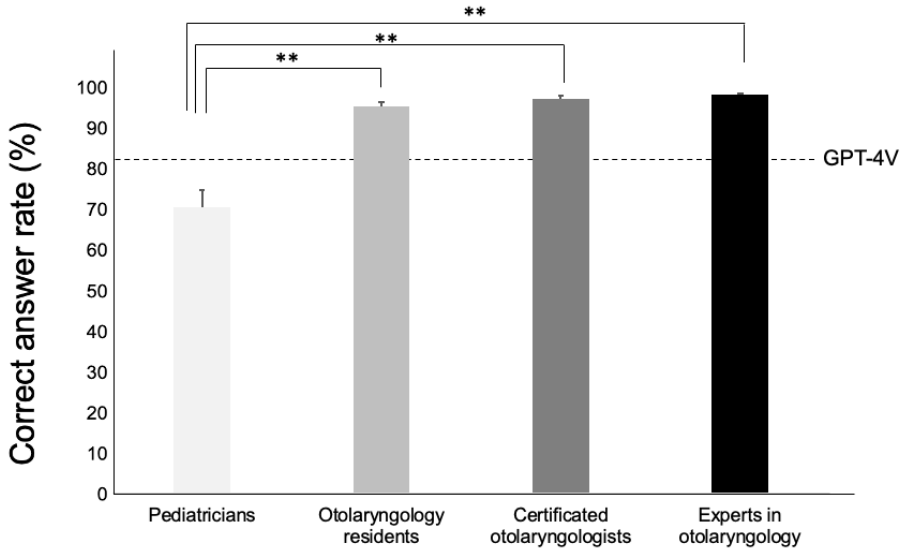
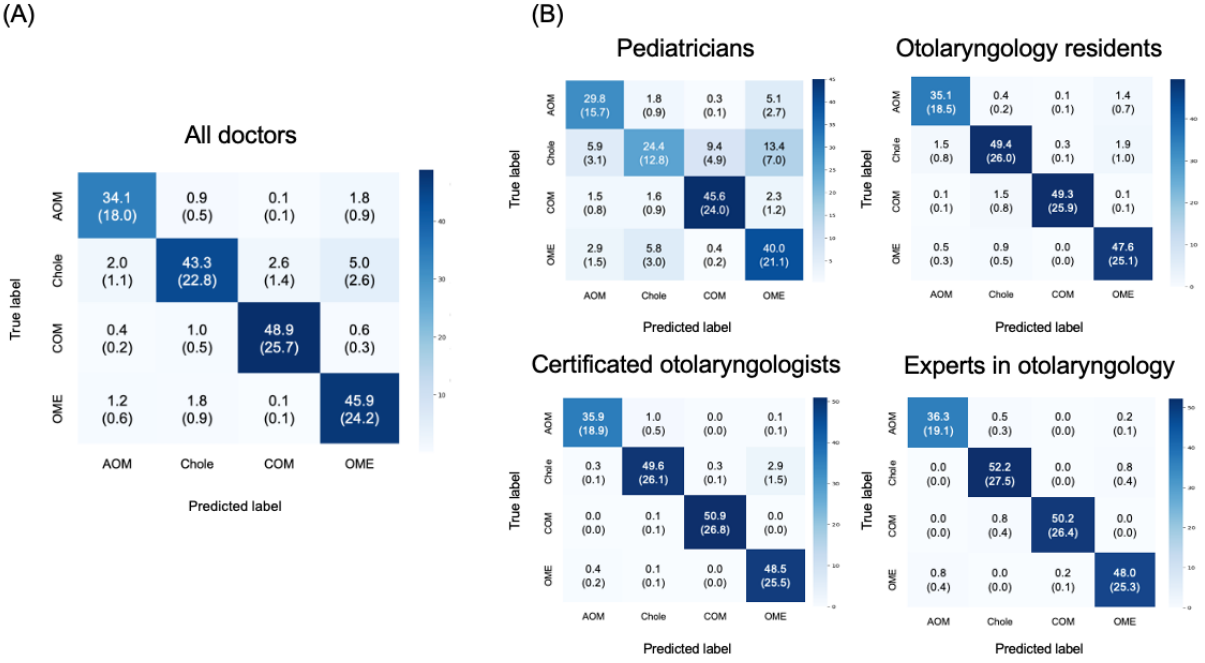


Figure 5. Confusion matrix of doctors (pediatricians, otolaryngology residents, certificated otolaryngologists, and experts in otolaryngology) for classifying 4 middle ear diseases. (A) Confusion matrix of all doctors (N=30). The average (percentage of total responses) is shown. (B) Confusion matrix 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]

[a]Statistically 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 36330706]
6. Alhudhaif 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 37779171]
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] [Medline: 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] [Medline: 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]

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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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]
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]
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]
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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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/lio2.1008] [Medline: 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] [Medline: 37318797]
27. Truhn D, Weber CD, Braun BJ, Bressem 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] [Medline: 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] [Medline: 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] [Medline: 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
PMID:38875669

Original Paper

Augmenting Telepostpartum Care With Vision-Based Detection of Breastfeeding-Related Conditions: Algorithm Development and Validation

Jessica De Souza[1], MSc; Varun Kumar Viswanath[1], MSc; Jessica Maria Echterhoff[2], MSc; Kristina Chamberlain[3], CNM, ARNP, IBCLC, MN; Edward Jay Wang[1], PhD

[1]Department of Electrical and Computer Engineering, University of California, San Diego, La Jolla, CA, United States

[2]Department of Computer Science and Engineering, University of California, San Diego, La Jolla, CA, United States

[3]Division 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

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

[a]FST: Fitzpatrick skin type.

[b]Not 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.

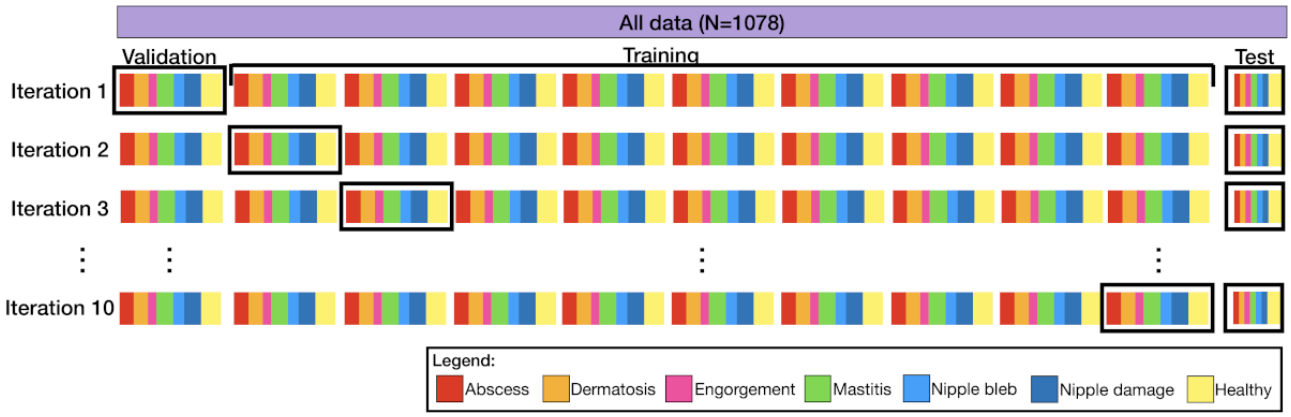
Figure 2. Samples of augmented data: (A) original, (B) brightness, (C) center zoom, (D) horizontal flip, (E) rotation, (F) shear, and (G) vertical flip.

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

[a]Unhealthy 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, F_1-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	*0.608* [a]	0.675	*0.623* [a]	*0.637* [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	*0.680* [a]	0.607	0.622
DenseNet169	*0.932* [a]	*0.771* [a]	0.507	0.659	0.596	0.572

[a]Italicized items represent the best metric.

[b]VGG16: 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.

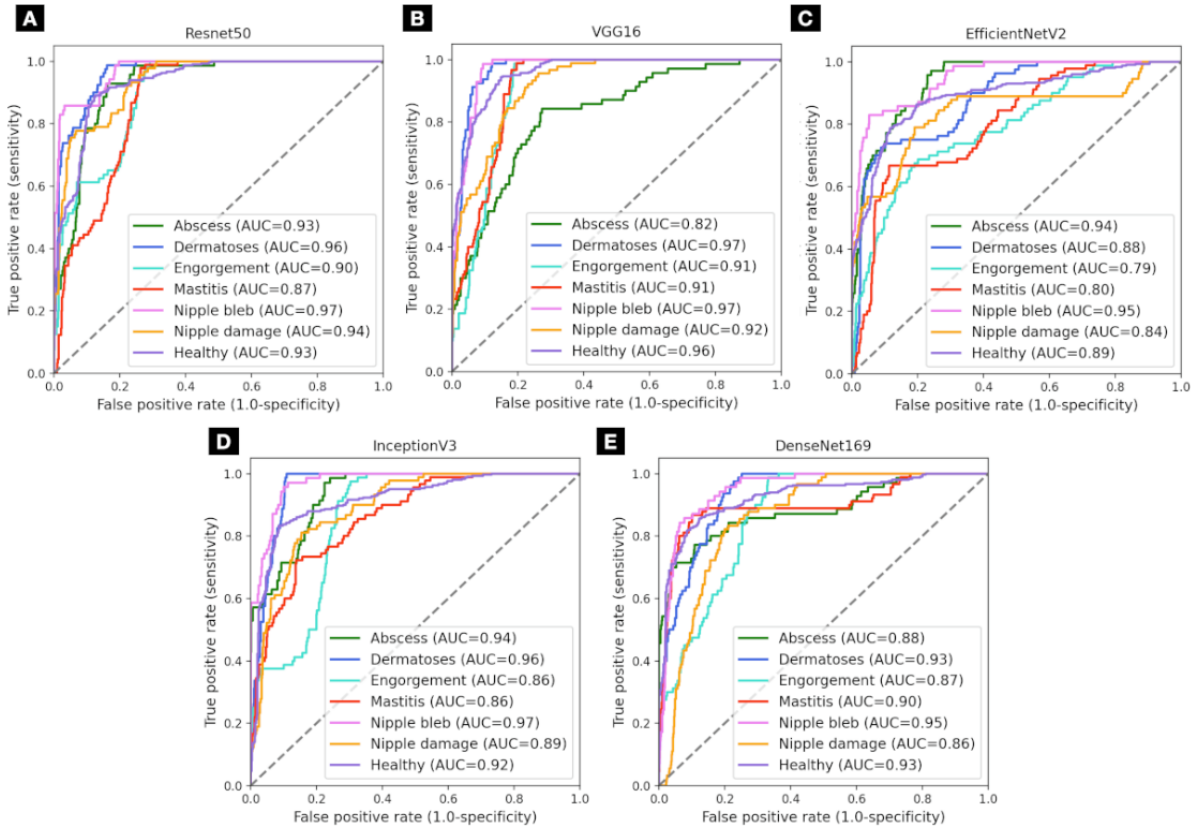


Figure 5. Performance of the 5 convolutional neural networks on the 7-class augmented data set: (A) Resnet50, (B) InceptionV3, (C) EfficientNetV2, (D) Visual Geometry Group model with 16 layers, (E) DenseNet169. AUC: area under the curve.

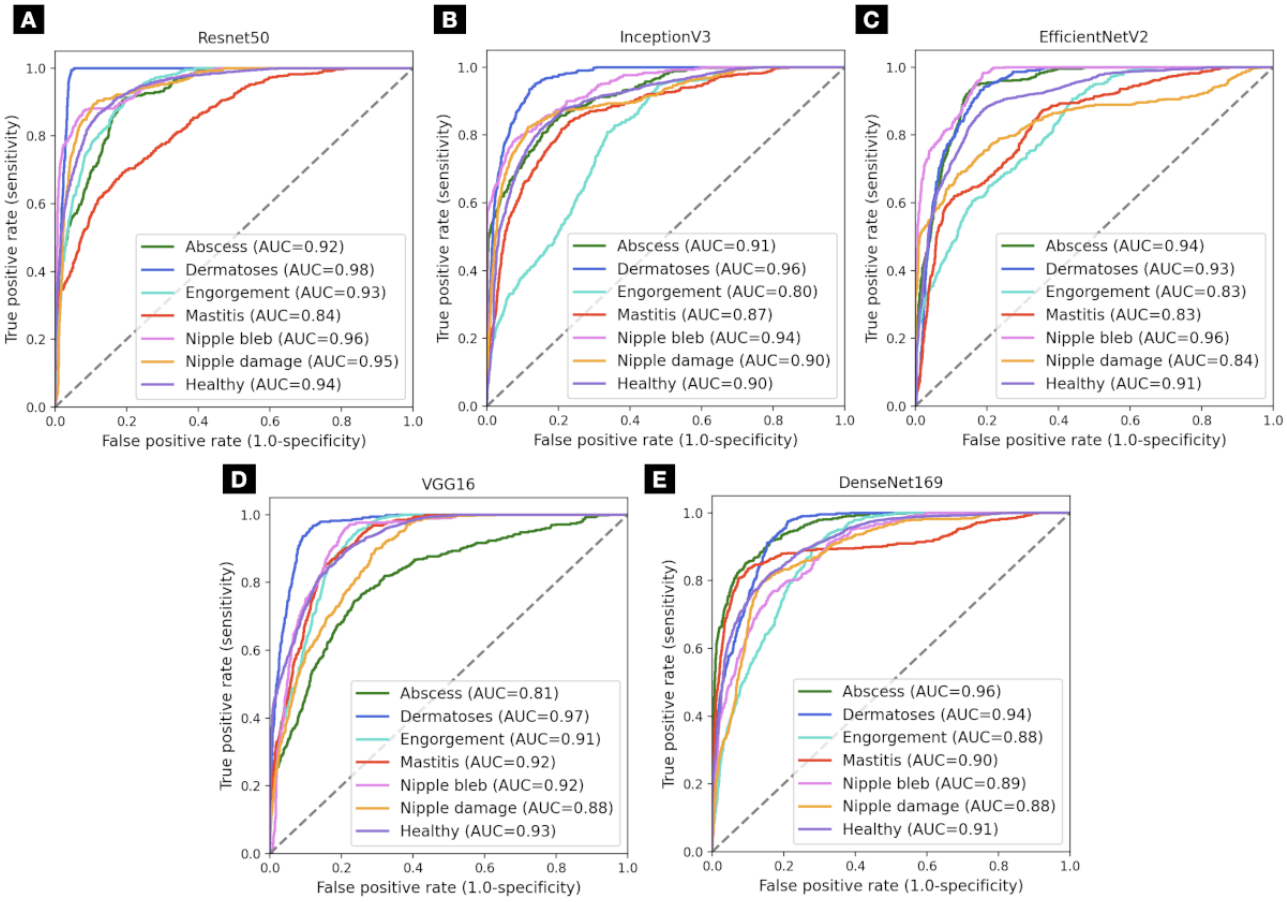


Table 4. Average evaluation metrics for the trained models on the augmented data set (sorted based on performance).

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	*0.907* [a]	*0.672* [a]	*0.717* [a]	*0.715* [a]	*0.713* [a]
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	*0.954* [a]	0.889	0.506	0.639	0.611	0.553

[a]Italicized items represent the best metric.

[b]VGG16: 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.

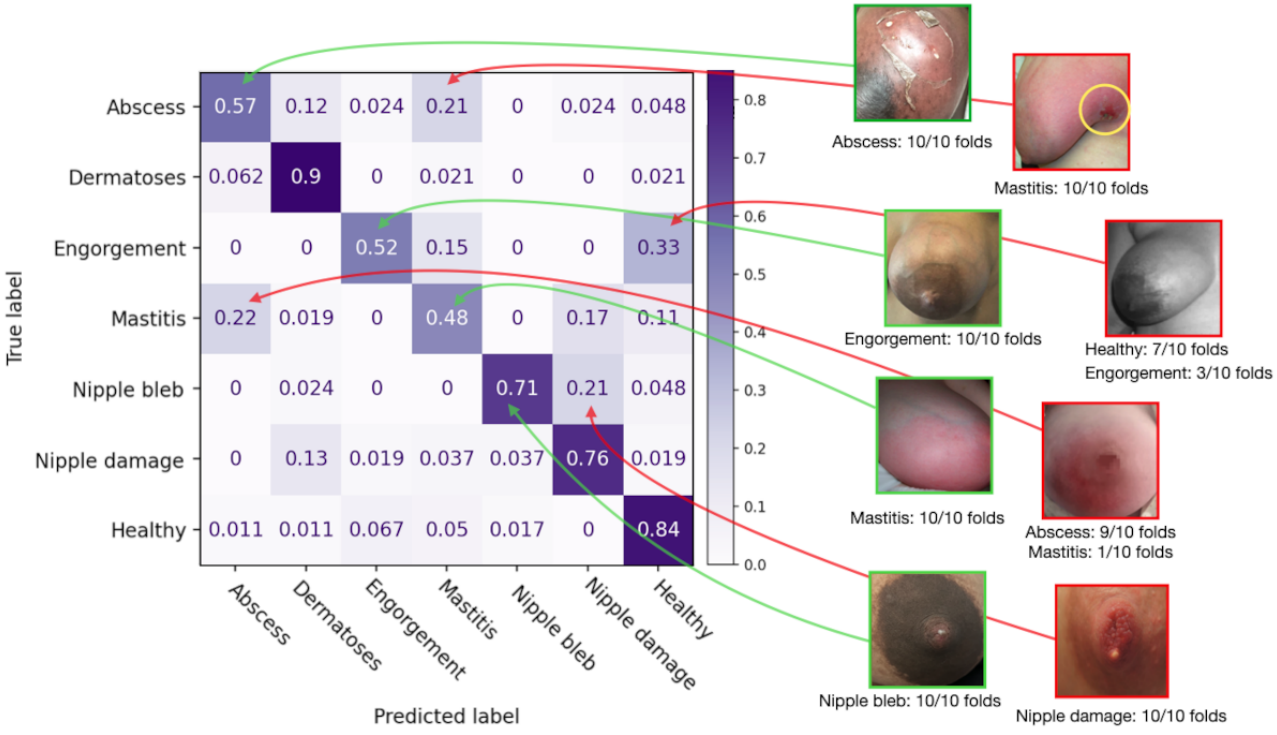


Table 6. Summary of the detection results per class: accuracy, precision, recall, F1-score, and support (ie, number of samples per class) using the Resnet50 architecture.

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	*0.877* [b]	0.990	*0.760* [b]	*0.859* [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	*0.991* [b]	0.629	0.769
DenseNet169	*0.935* [b]	*0.880* [b]	0.761	0.990	0.529	0.688

[a]VGG16: Visual Geometry Group model with 16 layers.

[b]Italicized 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.

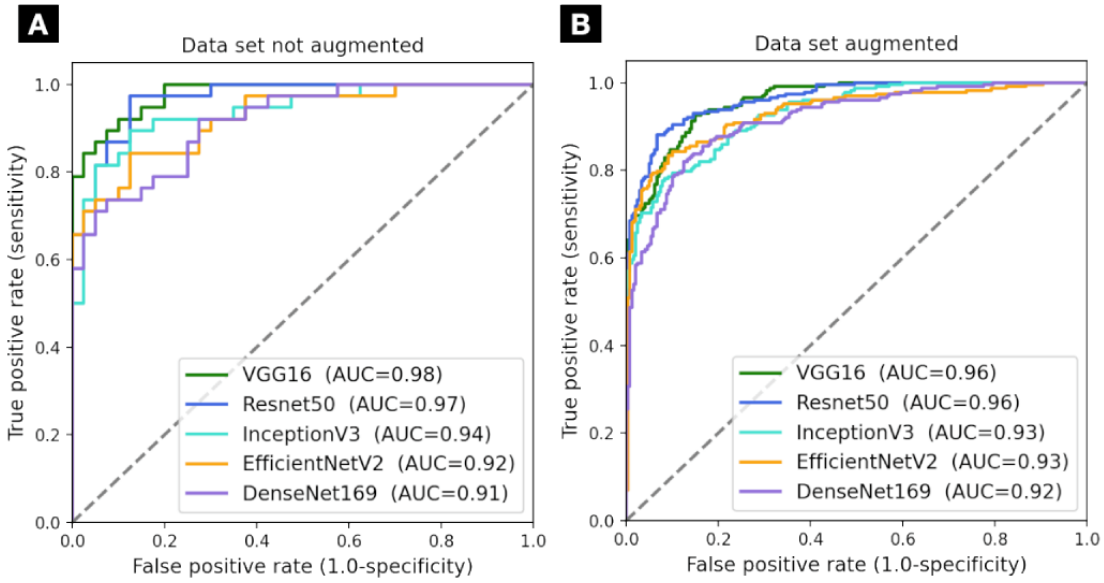


Table 8. Average evaluation metrics for the trained models on the augmented binary data set (sorted based on performance).

Data set and model	Training accuracy	Validation accuracy	Test set metrics			
			Accuracy	Precision	Recall	F_1-score
Binary augmented data set						
Resnet50	*0.952* [a]	*0.933* [a]	*0.877* [a]	*0.941* [a]	*0.801* [a]	*0.865* [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	*0.975* [a]	0.666	0.791
DenseNet169	0.946	0.927	0.771	0.952	0.570	0.713

[a]Italicized items represent the best metric.

[b]VGG16: 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.

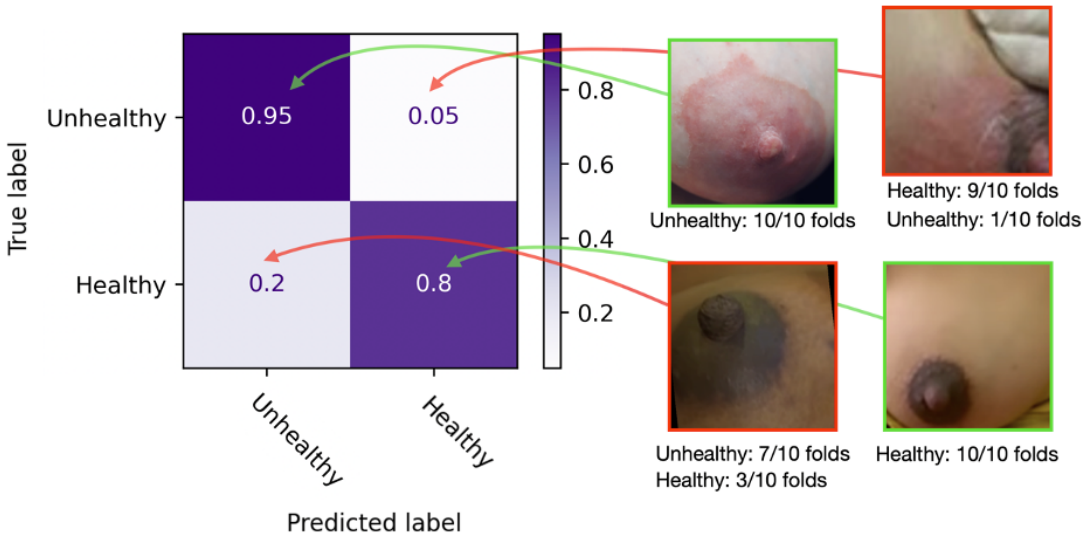


Image Quality

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.

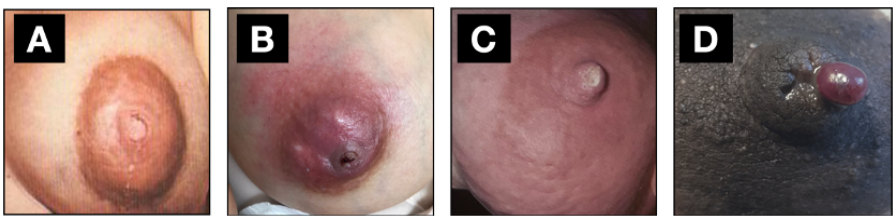
This highlights the need for (1) increasing the engorgement data set; (2) working closely with LCs to investigate the need to categorize conditions that can be a problem but indicate false positive cases of more serious issues; and (3) exploring the possibility of using these conditions that have higher 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 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.

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, 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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]
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] [Medline: 26644419]
13. Current statistics on worldwide IBCLCs. International Board of Lactation Consultant Examiners. URL: https://iblce.org/about-iblce/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] [Medline: 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] [Medline: 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]
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] [Medline: 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] [Medline: 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] [Medline: 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/fdgth.2023.1143528] [Medline: 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] [Medline: 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]
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] [Medline: 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] [Medline: 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] [Medline: 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]
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] [Medline: 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] [Medline: 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.0000000000000046] [Medline: 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]
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] [Medline: 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]
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] [Medline: 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]
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]
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]

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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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] [Medline: 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]
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] [Medline: 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] [Medline: 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]
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]
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] [Medline: 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]
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]
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]
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]
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]
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]
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] [Medline: 28241182]

66. Jain S, Singhania 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] [Medline: 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]
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]
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]
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]
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]
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]
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]
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] [Medline: 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]

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
PMID:

Corrigenda and Addenda

Correction: Using Conversational AI to Facilitate Mental Health Assessments and Improve Clinical Efficiency Within Psychotherapy Services: Real-World Observational Study

Max Rollwage[1], BSc, MSc, MPhil, PhD; Johanna Habicht[1], MSci; Keno Juechems[1], BSc, MSc, PhD; Ben Carrington[1], BSc; Sruthi Viswanathan[1], BTech, MRes; Mona Stylianou[2], BA, PGDip; Tobias U Hauser[1,3,4,5], PhD; Ross Harper[1], MA, MRes, PhD

[1]Limbic Limited, London, United Kingdom

[2]Everyturn Mental Health, Gosforth, United Kingdom

[3]Max Planck University College London Centre for Computational Psychiatry and Ageing Research, University College London, London, United Kingdom

[4]Department of Psychiatry and Psychotherapy, Medical School and University Hospital, Eberhard Karls University of Tübingen, Tubingen, Germany

[5]German Center for Mental Health (DZPG), Tubingen, 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

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
PMID:

Publisher:
JMIR Publications
130 Queens Quay East.
Toronto, ON, M5A 3Y5
Phone: (+1) 416-583-2040
Email: support@jmir.org